Robust, Efficient, Dynamic Theory of Mind

Marion Roth

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

June 2024

Abstract

Theory of Mind is a psychological term that describes the ability to reason about others' mental states. This includes their thoughts, feelings, beliefs, goals, or intentions of future actions. In reasoning about another's mental state, an individual can generate a vast number of possible theories about what is on their mind. In a rich and complex social situation, the options could be infinite. In theory, therefore, *mindreading* is a very sophisticated and energy-consuming ability. Computational simulations of detailed and sophisticated Theory of Mind require lengthy computations and elaborate processes. Yet, humans are able to reason about others' mental states with ease. Research has suggested that humans are not perfect at handling the contents of another's mind. Rather, human Theory of Mind shows errors and biases. This work argues for a conceptual approach to Theory of Mind as a responsive, selective, and incomplete cognitive ability. From a perspective of limited cognitive energy, it proposes that mindreading processes are shaped by individual and situational demands and resources.

This thesis is supported by four studies which explore what heuristics may be involved in reducing the cognitive energy an individual spends on Theory of Mind. The first study suggests ad hoc representations as a structural component reducing cognitive costs and identifies various inferential patterns that may narrow down the number of options to be considered in Theory of Mind reasoning. The second study explores stereotypes as a mindreading heuristic and considers the balancing act of navigating the own wants and needs, others' reactions, social belonging, and experiencing certainty. The third study investigates the phenomenon of egocentrism in more detail and the fourth study proposes that flexibility and gaps in mindreading can shape its efficient use.

The four studies are based on quantitative, qualitative, and mixed methods, and deliver insights into the complex and heuristic processes underlying Theory of Mind. Results are discussed with both psychological and computational considerations. Throughout this work, an overarching conceptual framework guides the exploration of different elements, within which various heuristics are identified and discussed. The thesis demonstrates a richness in Theory of Mind strategies that reflects large individual differences and imperfect yet robust abilities to reason about others. This work highlights the value of interdisciplinary work to generate novel perspectives and considerations and facilitate innovative research in both fields. Implications for future research and artificial intelligence applications are discussed.

Contents

Previous dissemination of findings

Chapter 5.1: Cutting Corners in Theory of Mind

Roth, M., Marsella, S., and Barsalou, L. (2022). Cutting Corners in Theory of Mind. In *AAAI Fall Symposium 2022 on Thinking Fast and Slow and Other Cognitive Theories in AI.*
URL: https://ceur-ws.org/Vol-3332/paper11.pdf

List of Tables

List of Figures

Acknowledgements

I am deeply grateful to my supervisor Stacy Marsella, for supporting me with my development and career since my undergraduate degree. He has always guided me towards the most fruitful means of solving problems within my own interests and ambitions. His own curiosity and passion for the field have very positively shaped my relationship with the topic and he has helped me expand my perspectives beyond what I ever expected. He truly sees the humans in his students and receiving his support throughout these years has been an incredible gift.

My gratitude also goes to my supervisor Larry Barsalou, who has never failed to share his experience in extraordinarily helpful ways. Providing me with so many resources and insights from years of being an outstanding researcher, it has been a privilege to receive his help and advice.

I also want to thank all members of the CESAR lab in Boston and Glasgow for their time to listen to my work in progress, and for sharing their own work with me. I am particularly grateful for being able to work with Nutchanon Yongsatianchot and Haley Matuszak, who have greatly enhanced the quality of my research by offering their expertise and ideas. I also immensely appreciate the support and input I have received by Tobias Thejll-Madsen, who has on many occasions been there to listen when I felt that my progress had stagnated, and with great patience, understanding, and new perspectives reminded me that it is worth it to hang in there.

Furthermore, I am very grateful for being able to collaborate with Jan Pöppel. Sharing his computational perspective on Theory of Mind with me has been very positively influential in shaping my research questions and developing the approaches to answer them.

Finally, thank you to the Social AI CDT and UKRI for the opportunity to work on this project and for the funding that made my research possible. It has been a truly exciting time being part of this interdisciplinary research group. Thank you to my many colleagues from different cohorts for their advice and feedback, and thank you very much to the fantastic team of administrators who have provided help and support throughout the years in so many different ways.

Declaration

All work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

Theory of Mind

Theory of Mind (from now on abbreviated to ToM), sometimes also referred to as *mindreading* or *mentalising*, is the ability to reason about others' mental states (Premack and Woodruff, 1978). For example, consider the scenario of a woman entering a busy cafe. When she approaches the coffee bar, she orders a cappuccino and the barista asks her for her name. She replies that she has a boyfriend, to which the barista responds "I mean for your coffee". In this scenario, both staff and customer have an understanding of what are expected and appropriate things to say and do, what the other person is likely going to say, and how their own behaviour will be received. Here, the woman misinterprets the barista's intentions, assuming that he showed romantic interest. He realises that she misunderstood him and elaborates on the reason for his question to clarify his intentions.

Figure 1.1: ToM example. Image from www.pikwizard.com

Mindreading involves perceiving and integrating communicative signals and cues and establishing representations of the contents of others' minds. This lets a person consider perspectives that are not their own. Humans commonly use this ability to understand others' feelings and behaviour and judge what actions will be appropriate in response (Baron-Cohen et al., 1985; Frith and Happé, 1994). In the barista example above, for example, better executed ToM could have helped the customer consider the barista's perspective and realise that he was asking for her name to process the order. Understanding and considering another person's mental states in social interaction is a fundamental component of human social interaction (Leslie, 1991). A person never has true access to another person's cognitive and emotional experience and yet, understanding and making sense of others' experiences is a central component of human social life (Herrmann et al., 2007).

This chapter is concerned with a general introduction of ToM, a review of relevant literature, and the definition of key concepts and terms. Subsequently, the rationale for this PhD will be outlined and specific research questions will be proposed.

1.1 Sally and Anne

The most commonly used paradigm to study ToM is the Sally-Anne test. It was developed by Baron-Cohen et al. (1985) to assess when children develop the ability to reason about others' mental states. Figure 1.2 illustrates the paradigm in detail. Sally places her toy in a box and then leaves the scene. While she is absent, Anne changes the location of the toy. When Sally returns to the scene, researchers would ask their participants where Sally would look for the toy. As Sally does not know that Anne changed the toy's location, she would falsely believe that it is in the original box. Typically, it is recorded if participants can come to this conclusion when they are asked to predict where Sally will look for the toy. The Sally-Anne paradigm captures differences in beliefs and explores how participants reason about them. Moreover, and more importantly, it demonstrates the representation of Sally's false belief, which is in contrast to the participant's knowledge of the true location of the toy.

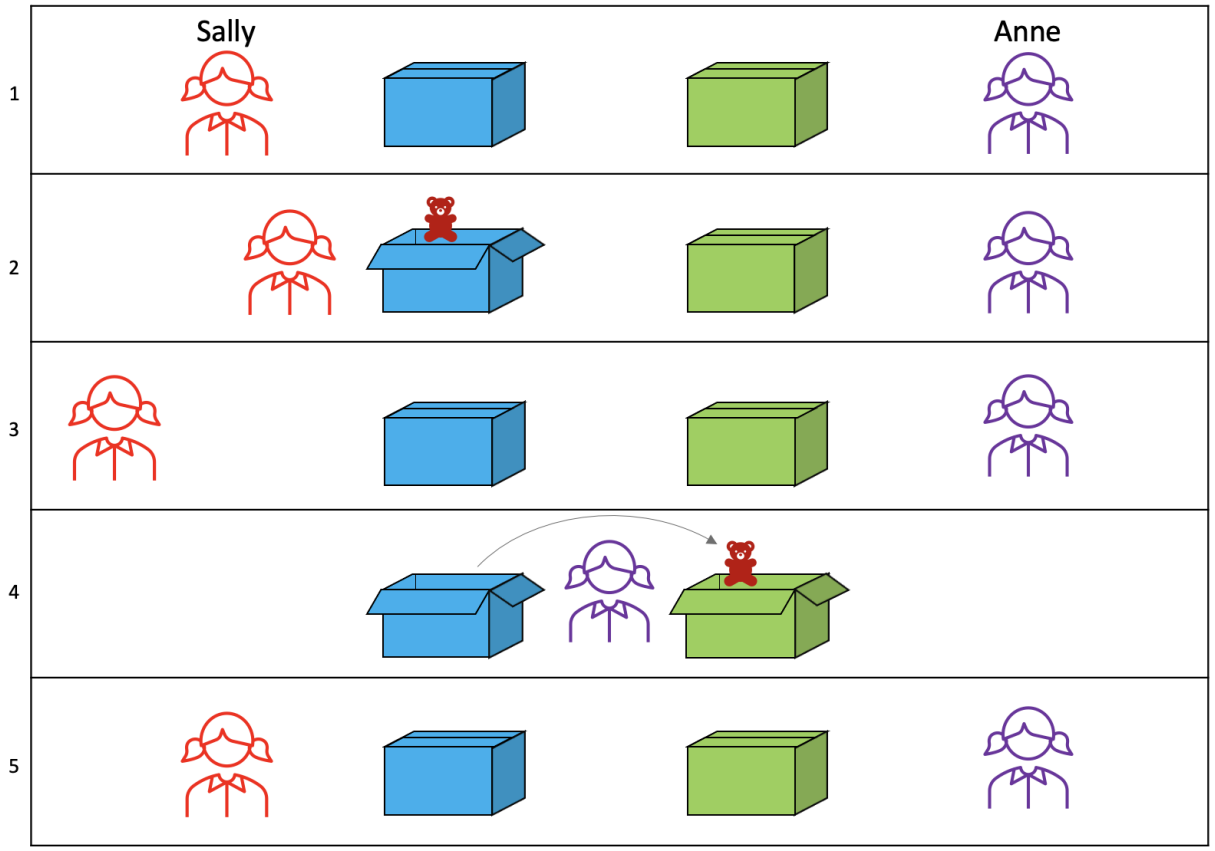


Figure 1.2: The Sally-Anne paradigm. 1. Sally and Anne are in the same room with two closed boxes. 2. Sally opens the blue box, puts a toy into the box, and closes it. 3. Sally goes for a walk and leaves the room. 4. Anne opens both boxes, moves the toy from the blue box to the green box, and closes both boxes. 5. Sally comes back. Where will Sally look for the toy?

Figure 1.3 illustrates the ToM phenomenon more generally. Another person's mental states, such as their beliefs, intentions, evaluations, emotions, or preferences, are fundamentally hidden and never truly accessible. Yet, they are a substantial component of human social interaction. Understanding and successfully navigating others' mental states is often fundamental to building relationships, establishing trust, learning, and managing everyday social spaces (Herrmann et al., 2007). The term "Theory of Mind" originally stems from the idea of establishing theories about what is on another person's mind (Premack and Woodruff, 1978). ToM research has been concerned with the cognitive processes that are required to establish such theories (Ong et al., 2019; Apperly, 2012; Gweon et al., 2011) , the scope and limits of ToM (Conway et al., 2019; Keysar et al., 2003), the development of the ability (Hughes et al.,

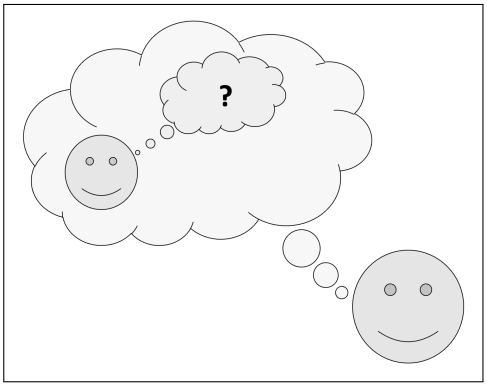


Figure 1.3: Illustrating ToM: Representing another person's mental states.

2005; Frith and Frith, 2003).

1.2 Definitions

Schaafsma et al. (2015) highlight the importance of operational definitions. Before exploring ToM in more depth and reviewing relevant literature, this section introduces definitions of the key terms and concepts this work refers to.

This work will commonly refer to *mental states*. These can include any emotional or intellectual experience by a person. Mental states are internal and distinct from but commonly connected to observable reactions and expressions (Thornton et al., 2020).

Mental models are the internal representation of an external reality (Jonker et al., 2010). Generally, the term describes the representation of any object or experience, including relevant knowledge, memories, evaluations, or expectations. In this work, the term will primarily be used to refer to mental models of other people, rather than objects. This includes information on others' behaviour, social roles, anticipated actions, attitudes towards them, and so on. ToM, in short, is the *mental modelling* of another's *mental state*.

The *Theory Theory* account characterises the first efforts to study ToM scientifically (Carruthers and Smith, 1996). Advocates for the approach suggest that we form models of others on the basis of generating inferences from observations. This perspective assumes interpretation and processing of our observations in order to create a model of them, which is then used to generate predictions of future behaviour. Thereby, within an already abstract space, we work backwards from a behaviour or expression to the underlying mental state.

Simulation Theory in contrast, proposes that person A's perceptions of person B's behaviour involves simulating B's behaviour as if it was A's own, for example using mirror neurons (Gallese and Goldman, 1998). Thus the same neural mechanism that A uses to reason about their own behaviour and act is repurposed to reason about B. A's model of B is grounded in this simulation rather than an inference from a theory about, or model of, B.

Explicit ToM describes the concrete and conscious reasoning about others' mental states. When performing ToM explicitly, individuals are able to express their beliefs about others' thoughts, feelings, intentions, or actions (Wimmer and Perner, 1983).

The term *Implicit ToM* has been used to describe a less thorough processing of others' mental states outside of the individual's awareness. Thereby they may cognitively integrate information about others' mental states, but not consciously so (Deschrijver et al., 2016).

Egocentric perspectives are primarily, if not fully, focussed on oneself rather than other people. In the context of ToM, it describes the instances in which an individual draws on their own experiences to reason about another person's behaviour or mental states. Southgate (2020) researches and discusses its opposite, *altercentrism*, which describes when a person is fully immersed and influenced by others' perspectives, rather than their own.

A *stereotype* is an example of heuristic-based processing, in which a complex concept or phenomenon is reduced to a much simpler cognitive representation. For example, all members of a social group may be represented as having the same limited set of attributes or characteristics (Madva and Brownstein, 2018).

Bayesian statistics is an approach to statistics that is based on probabilities. It has been used to model ToM by considering different degrees of certainty in others' beliefs (Baker et al., 2011).

Heuristics are generally understood as shortcuts that can lead to sufficiently accurate solutions to problems with significantly fewer resources. In Psychology, these shortcuts are typically associated with cognitive biases and generalisations to speed up perception, judgement, and decision making (Kahneman, 2011).

In economics, an opportunity cost calculation is the calculation of how to invest available resources in one alternative and thereby missing out on another (Spiller, 2011). Here this concept is adopted to refer to the investment of cognitive energy in ToM processes, depending on individual priorities and situational demands.

In the Cambridge Dictionary, a *robust* object or system is defined as "strong and unlikely to break or fail" (Peters, 2013). The dictionary defines *efficient* as "working or operating quickly and effectively in an organised way" and *dynamic* as "continuously changing or developing" (Peters, 2013). These features are considered here as characteristics of human ToM that requires a minimal level of cognitive resources while producing results of sufficient quality.

1.3 Literature

1.3.1 Psychology

Psychological research has explored the phenomenon of ToM over decades and from various angles. The following reviews some of the key insights from the work to date.

Theories

Theory Theory and Simulation Theory are the two traditional accounts of ToM. The theories differ in their perspective on whether forming mental models of others has its focus on top-down or bottom-up mechanisms. Over decades, these two main understandings of mindreading were competing with each other (Carruthers and Smith, 1996). The discussion is inherently linked with philosophical questions concerned with the inherent nature of reasoning, thinking, and beliefs.

Based on their research with primates, (Premack and Woodruff, 1978) initially coined the term *Theory of Mind*. They investigated whether chimpanzees could attribute beliefs and mental states to others and observed fundamental differences to humans, by comparing strategies in solving problems such as accessing food, fixing malfunctioning tools, or inferring an issue from

observing an actor's behaviour. Their research views ToM as the ability to use information from observed behaviour and expressions to infer the not directly accessible content of others' minds and generate predictions about their future actions. This approach is in line with the Theory Theory outlook on mindreading, highlighting the role of reasoning and understanding to learn about others' mental states. It emphasises the account of a 'folk psychology' perspective on reasoning, according to which we are motivated to make sense of others' behaviour in context, identify their intentions, and predict future actions (Ravenscroft, 2019). The Theory Theory account of mindreading emphasises the abstract nature of beliefs and their recursive characteristic: We can have beliefs about objects or people, but beyond this, we can have beliefs about beliefs, or beliefs about beliefs about beliefs. Jacob (2019) refers to these as *meta-representations* and explains their core role for ToM.

Theory Theory has particularly been challenged by evidence of mirror neurons and their role for cognitive processing (Preston and de Waal, 2001). Mirror neurons are active both when we perform and observe an action. Simulation Theory proponents argue that their activity causes a simulation of others' behaviours or feelings as we perceive them (Carruthers and Smith, 1996). The relevant evidence supports a focus on the role of empathy and perception for making sense of others' minds. Particularly, studies of individuals struggling to understand others' perspectives and having difficulties with mind-reading have drawn on the explanation of differences in mirror neuron activity compared to people showing typical patterns (Rizzolatti and Fabbri-Destro, 2010).

The debate between a simulation and a theory account has been maintained over several years, with evidence for both perspectives and not being accounted for by the other. For example, Jacob (2019) proposes an analogy in favour of the Theory Theory: It is possible to know that I like strawberries without in that moment experiencing or imagining the joy of eating strawberries. In contrast, there is a large body of evidence suggesting that we often act before or without at all being consciously aware of the action (Frith, 2013), let alone having theories about others which we can observe or recall.

Only in recent years, scientists have started to consider the two approaches together and even expand beyond it. Gallagher and Fiebich (2019) for example, introduced a pluralist perspective on mindreading. They propose that both simulations and inferences are common and useful tools for performing ToM. However, they point out that both come with the assumption of individuals as observers rather than the process of mindreading being creative and dynamic. They suggest that there are other and much more versatile tools beyond the two, which humans use to create, expand, or enrich our models of others, such as stereotyping, consideration of social norms and identities, direct perceptions, or embodiment cues. Furthermore, they argue that traditional accounts often neglect the role of embodiment and perceptual information informing ToM. Gallagher and Fiebich (2019) claim that mindreading is highly context dependent and that different strategies will be more or less useful than others in a given situation. Choosing the

least effortful strategy in a given moment, they highlight the role of the social, normative, and cultural framework within which it occurs. There is, however, still very little evidence on how exactly humans perform those skills they mention, and how they are dynamically combined.

Failing Theory of Mind

ToM as it is observed in adults is not present from birth. In fact, there is much debate and disagreement about whether it is innate or learned, that we can understand others as having beliefs that are distinct and may differ from our own and understand the content of those beliefs (Carruthers and Smith, 1996).

In their false belief task, Wimmer and Perner (1983) initially found that none of 3-4-year-olds, 57% of 4-6-year-olds, and 86% of 6-9-year-olds, answered the questions they were asked correctly by basing their response on the other person's false belief. However, more recent evidence has challenged those early findings. It has been argued that research may be confounded by researchers asking leading questions, children may want to please the experimenter, or not be capable of expressing their understanding, but nonetheless already use relevant information about others' mental states. By adapting the paradigm and recording looking time, Onishi and Baillargeon (2005) established non-verbal evidence of the ability to distinguish between true beliefs and false beliefs in 15-month-olds. Choi et al. (2018) even claim that their results show consideration of others' perspectives in children as young as 3 months old.

Southgate (2020) argues that even very young infants perform ToM. Pointing to evidence suggesting that already at a very young age children are capable of recognising and considering others' viewpoints and perspectives, she even suggests them to be altercentric, i.e. being very comprehensively aware of others' mind states, until their own self-awareness develops and conflicts with that ability. Burge (2018), in contrast, argues that there are very powerful other explanations for various findings which do not require mind-reading, such as learning based on associations or reinforcement. She raises the question whether conclusions about ToM abilities at such a young age are too far-fetched. Alternatively, it has been argued that young children are able to perform implicit, but not explicit ToM (Deschrijver et al., 2016).

The debate around ToM abilities and development has not yet been resolved. Interestingly, even much later in life, Theory of Mind performance is far from clear-cut. Keysar et al. (2003) present evidence of limited ToM in adults. Participants were able to correctly express others' beliefs verbally. However, in completing a task they were biased by their own knowledge despite the awareness that it was not helpful to answer the question (they knew that the person who they interacted with did not have this information). Furthermore, Conway and Bird (2018) present an account of degrees of ToM. They contrast the use of knowledge about false beliefs with that of true beliefs. Suggesting the potential of individual differences in ToM use in adults, they argue that ToM is not necessary for true-belief tasks, but if employed anyway, this may indicate a higher degree of ToM use in some people compared to others.

Avramides (2019) reviewed Dretske's perceptual account of the mind. The perceptual model suggests that we perceive others' minds much like we perceive objects. We can make direct observations about the world but also indirect ones, by gathering information from "reliable indicators". We can, for example, know that something is hot, not by touching it, but by observing the effect it has on other objects. Similar to Thagard's (2002) account of coherence, Dretske claims that our observations and how they fit together make us more or less confident in our beliefs. Avramides (2019) evaluates the perceptual account from the perspective of Cavellian doubt, which suggests that we often misunderstand others' minds, even when others are putting much effort into making themselves understood. She claims that the perceptual account ascribes more accuracy to humans' attribution of others' mental states than there really is and lacks in explanatory substance with regards to ToM. With the model not accounting for the range of mistakes we make, Avramides (2019) suggests that there must be a mechanism beyond perceptual knowledge. However, she does not dismiss the perceptual account as being wrong, but rather proposes that it is too simplistic to explain the variety of findings on mindreading.

Cognitive context

The understanding of the mechanisms involved in Theory of Mind sheds light on the nature of human cognition in general. Research findings, in turn, inform the methodologies of contemporary studies and our understanding of the ways in which beliefs are cognitively represented and how they can be measured and manipulated. Schaafsma et al. (2015) point to the interconnectivity of ToM with many other cognitive phenomena, such as emotions, distinction of beliefs, categorisation of stimuli, and conceptual knowledge. They suggest that the study of those individual cognitive elements can inform the overall understanding of ToM and the processes involved. The way in which more fundamental cognitive phenomena are conceptualised therefore shape the approach to embedded cognitive abilities such as ToM.

An interesting perspective on belief and knowledge representation, for example, which has been gaining popularity over time is that of network-like structures: the relationship between elements rather than the elements themselves capture and represent meaning. Thagard (2002) proposed a comprehensive account on coherence as a fundamental concept guiding many cognitive mechanisms and suggested offered various principles that characterise different types of coherence. Thagard proposes that beliefs manifest when they are coherent within the given context, and that we assess the reality of beliefs by judging them in relation to relevant observations, experiences, and evaluations. In the context of ToM, this account characterises beliefs as represented in and emerging from networks, which are activated during cognitive processing.

This understanding of cognitive organisation is similar to Clark's (2013, 2015) predictive processing conceptualisation of cognition. Like Thagard (2002), Clark proposes connections across various concepts and at different levels of specificity. His account proposes a hierarchical organisation of cortical processing that unifies perception and action in a constant circle of

predictions and their correction. Thereby, higher cortical levels feed back to lower levels to guide the consistent interpretation and integration of perceived stimuli. Applied to ToM, this account views mental models as hypotheses about others' mental states, which can be updated and refined when new observations create prediction errors (Kwisthout and Van Rooij, 2020).

Another more general account of cognition which has been applied to ToM is that of a dual-system architecture. Dual-process perspectives are popular across various areas of psychology (Chaiken and Trope, 1999; Kahneman, 2011). Generally, thereby, it is distinguished 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. Representations are explicit and accessible, which allows control over processes, but comes at the cost of high mental effort. Dual-system approaches hold interesting implications for the process of learning: Novel tasks are processed more slowly and deliberately, and as we acquire skills, they take less time and cognitive resources. However, research suggests that, if there were two systems, the relationship between those different mechanisms would be much more complex than that. If there are two fundamentally different processes, they are highly interactive and complementary (Jacob, 2019).

The two-system account of mindreading aims to resolve the observed conflict between flexibility and efficiency in ToM and also aims to understand the inconsistencies in the evidence at hand (see Jacob, 2019 for an overview). It has been put forward to explain the discrepancy between evidence suggesting a lack in mindreading in children up to a certain age (Wimmer and Perner, 1983) on one hand and the evidence in favour of ToM already in very young infants (Onishi and Baillargeon, 2005) on the other hand. Apperly and Butterfill (2009) strongly argue in favour of a dual-system account of mindreading, suggesting that the fast and implicit component of ToM 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 (Jacob, 2019). Some ToM research focusses primarily on explicit ToM and some primarily on implicit ToM, whereas other researchers discount the distinction altogether (Burge, 2018).

On a lower level, the structure, representation, and adjustment of beliefs characterise the space in which we can perform ToM. The discussion of how beliefs are represented, structured, and maintained is applicable to the literature on categories. The notion of categorisation is based on ancient philosophical ideas about how we make sense of the world (Medin and Smith, 1984). Concept and category formation as conceptualised by Piaget (1976) draws on two main principles: assimilation (recognising a perceived stimulus as an element of an established category) and accommodation (forming a new category when a stimulus does not fit with any of the already existing categories). Current research questions around categorisation are concerned with their

representation, their level of specificity, and their role in relation to other cognitive processes. This includes the phenomenon of abstraction as a fundamental aspect to shifting across different degrees of generality. Barsalou (2003) discusses abstraction in perceptual symbol systems as a theory that captures both dynamic and sufficiently structured abstraction processes.

Specifying a particular instance of categorisation, Barsalou (e.g. 2019) has studied ad hoc categories. He proposes that we very commonly create these categories on the spot, to flexibly process the problems we solve. Rather than being fixed and firmly established, ad hoc categories are goal-oriented, formed flexibly, and context-dependent. When created and used repeatedly, an ad hoc category will turn into a more established and more easily retrievable belief structure.

Beyond this, there is much consensus today that cognition, reasoning, and consciousness are inherently connected with the way in which our brains are organised, and how electrical and chemical signals move within them, shaped by experiences of and within the physical world (Clark, 2013). An embodied cognition perspective on the present issues would not only consider the connection between body and mind but view it as the fundamental mechanism driving our thoughts. Bianco (2022), for example, argues that sensorimotor abilities and embodiment fundamentally contribute to and shape mentalising abilities. Barsalou's (2019) proposal of situated action suggests that rather than cognitive mechanisms occurring in isolation, they inherently emerge from the connectivity between the components involved. It highlights the inter-connectivity of brain, body, and environment: the concepts and beliefs stored in our brains need to be retrieved and used in vastly varying situations and are therefore required to be highly dynamic, adaptable, and compatible with the format of perceptions of the physical world. Badcock et al. (2019) highlight that a system's successful operation within and interaction with its environment requires appropriate structures to process it. This results in the observation that the system's architecture recapitulates the structure of the environment within which it exists (Badcock et al., 2019). Adopting this perspective suggests that our experience of others' knowledge and beliefs is shaped by the boundaries of our senses and how those determine the ways in which we can processes others' beliefs.

Furthermore, there is evidence that human cognition follows the principles of bounded rationality (Simon, 1956; Lieder and Griffiths, 2020): The resources available to an individual determine the bounds in which decisions and actions are rational or not. This perspective considers that humans do not have unlimited internal and external resources and proposes that behaviour is rational if it is worth the expenditure of the required resources in whatever shape or form. Heuristics are a cognitive tool that shape this phenomenon. A heuristic is a cognitive shortcut that reduces the resources required but allows the individual to still achieve a sufficient result. Their role in human cognition is increasingly recognised in Psychology and Cognitive Science (Kahneman, 2011). The role of heuristics in ToM, however, is less explored than in other areas of Psychology.

Defining Theory of Mind

Over decades of research, ToM has been investigated from many different angles and with various methods. The various perspectives on ToM and conflicting pieces of evidence for various aspects of ToM rest on very different understandings of what ToM is. Schaafsma et al. (2015) highlight the increase in studies concerned with ToM especially in the last 20 years. They claim that the way in which the ability is investigated is very rarely based on clear definitions of the concept and their overlap across studies. Furthermore, researchers utilise different measurement tools and paradigms, such as explicitly assessing attributions or judgements (Newen and Wolf, 2020; Bardi et al., 2016; Wimmer and Perner, 1983), observing behaviour in strategic games (Goodie et al., 2012), or tracking where a person looks or how task performance may be facilitated (Deschrijver et al., 2016; Freundlieb et al., 2015; Onishi and Baillargeon, 2005). Schaafsma et al. (2015) call for a systematic deconstruction and reconstruction of relevant terms, concepts, and associated appropriate measures, to achieve transparency and comparability of different research findings.

Theoretical perspectives and experimental findings in the field of ToM vary widely. A dual-process perspective, for example, encompasses very different mental capacities compared to a Theory Theory perspective or a perceptual account. This work restricts itself to a cognitive perspective, which is concerned with the inferences of others' mental states, underlying and explaining their behaviour (Csibra, 2017). This work explores the cognitive processes involved in making inferences about others and the costs associated with those processes. Thereby it adopts a Cognitive Science perspective in order to facilitate a more specific approach to identifying fundamental elements and computational mechanisms involved in ToM.

The following section reviews the Artificial Intelligence (AI) and Cognitive Science literature on Theory of Mind. Following on the review of previous research, we will outline the specific cognitive processes and their associated cognitive costs that are investigated in this work, which the rest of this thesis will be based on.

1.3.2 AI and Cognitive Science

ToM research has its origin in Psychology but the aim to understand mental state representations has progressively crossed over into computational applications.

Interest in computational Theory of Mind

Understanding Theory of Mind has become increasingly interesting to researchers in social robotics and AI (Yang et al., 2018). There has been increasing interest in modelling ToM computationally and the implementation of typical human traits in AI is exploding as a research field. Firstly, the AI industry is increasingly interested in including mental models in virtual agents to improve their effectiveness in various areas of application (Lake et al., 2017). Secondly,

computational models are a useful and popular research tool to better study human cognition and behaviour, even more so the more realistic and accurate they are. Thirdly, in the process of developing AI that smoothly fits into human society, the demand for social intelligence in artificial agent grows (Yang et al., 2018). Human-AI team performance, for example, has been shown to be influenced not just by AI accuracy but also humans' mental models of AI, and vice versa (Bansal et al., 2019). Benefits of improved AI involving ToM abilities can manifest across a range of applications, such as industry (Rocha et al.,), health (Garcia-Lopez), teamwork (Bansal et al., 2019), or safer interaction of robots with humans (Blum et al., 2018).

Modelling ToM

Traditionally, many attempts of modelling human-like computational agents were based on a BDI architecture: Bratman's (1987) theory of intention conceptualises the way in which we work towards achieving goals, and distinguishes between beliefs, desires, and intentions (BDI). Thereby, goals are desires of longer-term outcomes. These are broken down into much smaller steps, concrete intentions, which are associated with the execution of specific actions. The process of planning and executing is monitored by and filtered through the constant update of beliefs, based on what the agent observes. However, Herzig et al. (2017) point out that traditional models employing this approach do not account for higher-order beliefs, i.e. beliefs about others' beliefs.

A key characteristic of ToM reasoning is what Jacob (2019) refers to as meta-representation: our beliefs about (our or others) beliefs, the mental representation of a mental representation. Meta-representation requires recursion: we have beliefs about others' beliefs about our beliefs, and so on. Gmytrasiewicz et al. (1991) applied a recursive modelling approach to the Prisoner's dilemma, modelling an agent's decisions which considered hypotheses about the other agent's strategies. By means of decision trees and matrices, agent 1 generated hypotheses about the strategies of agent 2, its hypotheses about agent 1, and so on. Interestingly, they illustrate that at a certain depth, going deeper on the levels of recursion adds very little new or useful information. Importantly, the generation of hypotheses about another person's mental states could be infinite in the complex reality humans live in. Considering and reasoning about all possibilities of another person's mental states would be an impossible cognitive reasoning task.

The Nengo-based Semantic Pointer Architecture by Blouw et al. (2016) aims to model the biological structure of the brain. Neuron-like structures, representing elements and connections between them, pass on signals that can be modified and combined in various ways at each point of transmission. The so-called semantic pointers are organised as a network and connect different elements, i.e. from one element they point to other associated elements. Together they make up the representations of semantic concepts. These binding features allow for the representation of both basic and more complex concepts. The architecture at hand is very applicable to many areas of cognition, tapping into the issue of how concepts are structured, how we can adjust our

focus of specificity, and how information can be abstracted. Kajić et al. (2019) have used this idea and architecture to surprisingly accurately model emotions in response to different stimuli, even outperforming Eliasmith's Spaun (2012).

With growing insight that the brain operates like a "prediction machine" (Clark, 2013), researchers have increasingly used Bayesian approaches to model Theory of Mind. Computationally, the probabilistic element allows for more flexible and dynamic ToM reasoning. Baker et al. (2011) proposed a Bayesian Theory of Mind (BToM) framework, which is characterised by inverse planning. Hidden mental states, representing desires and beliefs, cause an agent's behaviour. The model probabilistically infers desires and beliefs from observations of the environment.

Alfonso et al. (2015) have proposed a probabilistic model integrating predictions (top-down) and perceptions (bottom-up). The agent employs a reverse engineering approach to infer beliefs about what is being observed, and additionally updates existing beliefs with newly observed information. Similarly, Yongsatianchot and Marsella (2016) modelled the integration of perceptions (bottom-up) with appraisals (top-down).

PsychSim (Marsella et al., 2004) is a modelling framework which offers both recursion and Bayesian predictions: The virtual agent can generate models of others, which include their beliefs about others, including their beliefs, and so on. The system uses the integration of predictions and observations to generate the inferences that are most likely in a given situation, and recommends actions based on them. However, the simulations are very costly, given that they are so rich and detailed, and are likely still missing elements to achieve even more realistic modelling of human reasoning, decision-making, and behaviour.

Pöppel and Kopp (2019) have developed a switching mechanism, which aims to resolve the issue of costly simulations. Their approach is based on the strategy of using simpler models for as long as they work and employing more complex and detailed models only when the simple models are not sufficient anymore.

Models vs humans

Even the probabilistic models of ToM, which perform closest to human ToM (Pöppel, 2023), have an intrinsic shortcoming: Human cognition operates fundamentally differently. Bayesian approaches use inferences based on reversed engineering: A pattern of mental state and behaviour is observed, stored, and used to predict the underlying mental state of a newly observed behaviour. The brain employs similar mechanisms of neural prediction and their refinement by means of prediction errors (Clark, 2015). However, even though the Bayesian approach (Baker et al., 2011) is fundamentally aligned with what research has shown about the predictive nature of the human brain (Clark, 2015), it does not consider the phenomenon of heuristics to drastically reduce cognitive costs. Beyond the ability to generate inferences, the brain is characterised by a complex organisation of different brain areas, a lifetime of developing cognitive and so-

cial strategies, and a limited availability of resources. Kwisthout and Van Rooij (2020) outline tractability problems with Bayesian approaches to modelling human cognition.

This work studies the processes underlying ToM to find more explicit shortcuts and mechanistic heuristics. It explores the bounds on cognitive processes that produce the cognitive inferences at the heart of Bayesian approaches. Thereby, it considers a perspective of opportunity costs to conceptualise the trade-off between different ToM features that may compete over available cognitive resources. In line with a bounded rationality perspective (Simon, 1956; Lieder and Griffiths, 2020), opportunity costs are understood as mutually exclusive desirable outcomes which cannot all be invested in with the available resources. This thesis employs a computational angle on the investigated psychological phenomena to facilitate the understanding of the cognitive processes involved. The computational angle forces the identification of clear processes and a more explicit investigation of their cognitive costs. In the following, the processes involved in ToM as decided on by the researchers for this work will be specified. Following, the cognitive costs associated with these processes will be outlined.

Cognitive processes

As mentioned above, this work takes a cognitive approach to ToM. This defines ToM as an inference process which links others' behaviours to underlying mental states (Csibra, 2017). By this definition, therefore, ToM includes the attention to and observation of a behaviour, forming a representation of the other's mental state, i.e. a mental model, and a set of beliefs about which mental states are associated with what behaviours. The representation of the other's mental state is used to guide decisions about future actions.

Relating this to the cafe example at the beginning at the chapter, the barista and the customer both observe each other and pay attention to specific aspects of their experience. The barista may be focussed on making a good coffee and following the correct procedures required for his work. The customer may be conscious of her interaction with a male stranger in a public place. A mental representation of the other person is formed by considering the current experience and previously stored mental model about what to expect and how to act from a person in this role and in this type of situation. Here, the customer may consider her mental model of male strangers, rather than the mental model of a barista, including the relevance of the name for her order. The barista may remember the importance of asking for customers' names on a busy day to get the right coffee to the right person. The mental model, in turn, can inform stored beliefs and understandings of the social world. Finally, the representation of the other person's mental states affect a person's decisions and thereby actions.

The barista is aware of a busy cafe and knows that organising coffee with customers' is a helpful strategy to get all orders right. He knows that efficient and reliable service will increase the chances of gaining a returning customer, and asks for her name. The customer, on the other hand, is conscious of her interaction with a male stranger in public. She considers her previous

experiences and infers that the barista may have romantic intentions. She decides to let him know that she is not available. The barista considers this response and his own knowledge about possible intentions around asking for strangers' names. He concludes that she misunderstood his intentions and may alter his beliefs about how a customer may perceive the question and how it could be adjusted in the future. He decides to clarify his intention by explaining that he needs her name for the order, so that she can adjust what mental model she uses to accesses to react appropriately.

Each elements in this process has a cost of cognitive energy attached to it. Computation over a large number of beliefs is exponentially more costly. A perspective of cognitive costs associated with ToM is considered in the following section.

Cognitive costs

Even though an individual may not always perceive it as such, cognitive processing, including ToM, is inherently effortful. Like with other cognitive phenomena, the processes involved in ToM will require the expenditure of cognitive resources. Effectively observing another person, for example, involves an effort of perceiving their actions so that these perceptions can be processed and stored. Similarly, the retrieval and consideration of already existing memories takes cognitive processing and energy. Inferences made about another person then require computation (Baker et al., 2011). Additionally, beyond the inferences themselves an individual will consider situational goals, resources, rewards, and other factors to produce inferences of different degree, type, or nature (Conway and Bird, 2018; Keysar et al., 2003), and determine whether it is worth the effort to generate an inference in the first place. Finally, for ToM reasoning to affect subsequent decisions and actions, the results need to be passed on to other cognitive structures, which also requires energy.

Humans show a fundamental bias towards less cognitive effort (Fiebich and Coltheart, 2015). ToM therefore needs to be managed in consideration of other situational goals and demands. If cognitive energy is conceptualised as a limited resource or currency, it can only be spent on a limited amount of costly processes, here ToM processes. Different costs and benefits in this opportunity cost calculation may be prioritised, but there is an overall limited amount of cognitive energy available to spend. Therefore, by pursuing one alternative, an individual will always miss out on another. For example, a person's attention is limited (Driver, 2001) and needs to be allocated, i.e. the attention required for this process cannot be spent elsewhere. Sometimes, however, it is very important for an individual to invest in the accuracy of their inferences, and may need to sacrifice attention to other stimuli in the environment. Furthermore, there can be social costs attached to certain inferences. For example, finding out about undesirable traits of a social in-group member may cause internal conflict and take up more energy (Matz and Wood, 2005; Sande and Zanna, 1987). Additionally, the integration of newly established inferences with other knowledge may require re-structuring of the existing belief network. Cognitive re-

structuring or consolidation of knowledge requires energy (Tononi and Cirelli, 2014). This trade-off needs to be managed cognitively, which also requires cognitive resources (Fechner et al., 2018; Matsuka and Corter, 2008; Coull, 2004). There are different theories on what constitutes cognitive energy expenditure, such as metabolic waste, employed resources, the type and number of decisions to be made (Westbrook and Braver, 2015) or perceived effort (Dunn et al., 2019; Garbarino and Edell, 1997). This thesis primarily considers cognitive effort and perceived emotional discomfort as indicators of spent cognitive energy in qualitative analyses.

1.4 Outstanding questions and issues

Over the years, different competing theories about ToM have been proposed and investigated. Increasingly, there is consensus that none of them hold in isolation, and true human ToM is likely a complex combination of various different elements (Gallagher and Fiebich, 2019). Moreover, ToM is not a fully reliable ability, as even adults often show gaps in their reasoning about others (Keysar et al., 2003). Researchers have considered more general insights about human cognition to understand the mind-reading ability, such as predictive processing patterns (Clark, 2015), embodied cognition (Bianco, 2022), or dual processing accounts (Jacob, 2019). The present work also considers accounts of coherence (Thagard, 2002), belief structures (Badcock et al., 2019), and categorisation (Barsalou, 2019), to drive a better understanding of ToM.

It has been highlighted that much of ToM research is based on very different understandings of ToM (Schaafsma et al., 2015). Here, ToM is conceptualised from a cognitive perspective with focus on the processes underlying the inferences about others' mental states. Thereby, we particularly consider the insights on bounded rationality (Simon, 1956) and further investigate the role of heuristics in ToM (Lieder and Griffiths, 2020; Kahneman, 2011). Current theories of human Theory of Mind are still predominantly abstract and do not provide the in-depth understanding required to realise a concrete mind-reading model (Avramides and Parrott, 2019). The aim of this work is to identify more concrete cognitive processes associated with ToM, their associated costs, and the heuristics that can keep these costs low. This is to form a basis for eventual modelling that is more aligned with human cognition and behaviour than current models are.

Computational perspectives on ToM fundamentally differ from psychological perspectives. Approaches to computational modelling of mind-reading (Bratman, 1987; Gmytrasiewicz et al., 1991; Pynadath and Marsella, 2005; Baker et al., 2011; Eliasmith et al., 2012; Alfonso et al., 2015; Kajić et al., 2019; Pöppel and Kopp, 2019) are typically aiming to reproduce human performance. However, an incomplete understanding of the underlying mechanisms challenges researchers in developing architectures and mimicking human processes in computational manifestations of ToM. Rather, even well-performing ToM models have been proposed to be intractable, due to overlooking the biases and heuristics employed in human ToM (Kwisthout and

Van Rooij, 2020; Lieder and Griffiths, 2020). In contrast to humans, who can learn from only very few samples (Banich and Caccamise, 2011) many computational models require large sets of learning data (Gandhi et al., 2023).

Bayesian approaches aim to align with the human brain's predictive processing characteristics (Clark, 2015) but do not consider the phenomenon of heuristics and structural factors that lead to drastic reduction of cognitive costs. In this work, bounded rationality, heuristics, and situational demands are key concepts that shape the perspective on both psychological observations and computational interpretations. Ultimately, a person has only so much capacity to perform a task as complex and sophisticated as reasoning about other's mental states. There are mechanisms to be identified and specified that allow humans to manage their rich social worlds with the resources they have available. These phenomena have been considered theoretically (Kwisthout and Van Rooij, 2020; Gallagher and Fiebich, 2019) but not practically, in a modelling context. To achieve human-like ToM performance, a computational model of ToM would ideally reflect the mechanisms that drive human ToM. However, as reviewed above, human ToM is still not understood in sufficient detail to date, to develop such a model. This work aims to generate experimental ToM research that considers modelling perspectives and thereby propose requirements to bridge the gap between the two fields. Based on these outstanding issues and questions, the following overall research question guides this thesis.

Research Question:

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

Schaafsma et al. (2015) discuss the increased use of the term 'Theory of Mind' in many different and conflicting ways. They call for the deconstruction and reconstruction of how the concept is understood and suggest which various cognitive elements may be involved in ToM. They suggest the following elements to be fundamental components of ToM and recommend their study to untangle the various insights and understandings of the concept: "perceptual discrimination and categorization of the socially relevant stimuli, as well as of interoceptive signals elicited by those stimuli, semantic or conceptual knowledge, executive processes, and motivational processes" (Schaafsma et al., 2015, p. 68). There may be more relevant elements than that but this thesis adopts these elements as a guiding thread in the study of heuristics in ToM. Additionally, this work considers the component of mental state representations.

This thesis aims to answer the research question stated above by considering which heuristics may characterise those different elements. Figure 1.4 visualises all elements considered in this work and will be shown in different chapters to illustrate where the respective heuristics fit into the overall picture of ToM.

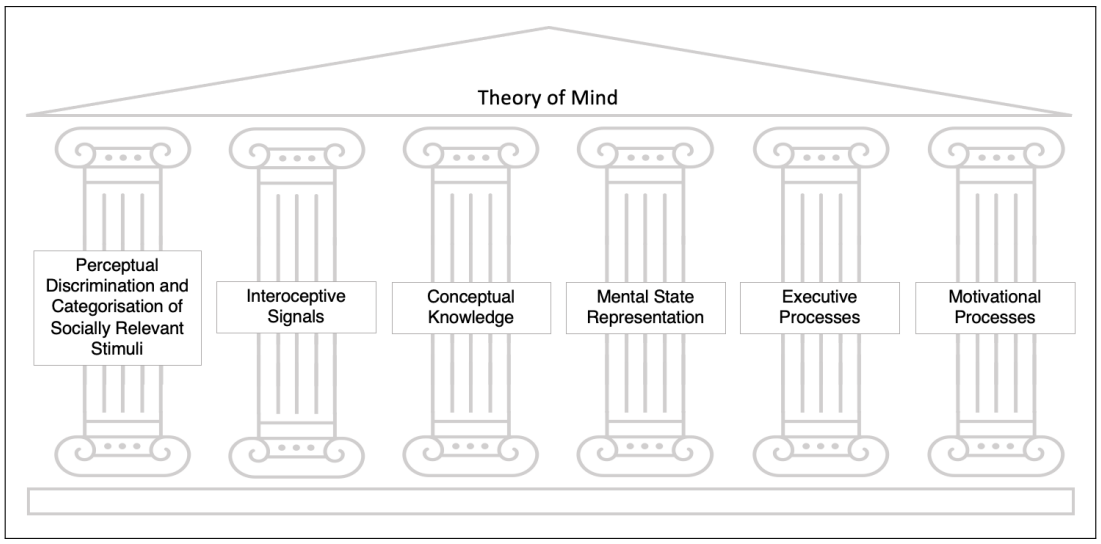


Figure 1.4: ToM elements, based on Schaafsma et al. (2015). "Perceptual discrimination and categorization of the socially relevant stimuli, as well as of interoceptive signals elicited by those stimuli, semantic or conceptual knowledge, executive processes, and motivational processes" (Schaafsma et al., 2015, p. 68) are examples of elements to study. Here the element of mental state representation itself is added.

This work aims to establish evidence and insights that are based on the study of human behaviour and experience but are informed by modelling perspectives so as to eventually be useful for and compatible with the generation of concrete models. Particularly, the relationship and trade-off between different cognitive resources and costs will be investigated to shed light on mechanisms in place that help to provide for robust and dynamic, but yet efficient ToM in human behaviour. Both psychological and computational perspectives are considered to drive a more comprehensive understanding of ToM processing and practical modelling outlooks. The focus is kept on the underlying structure and processes of human ToM, and what concrete heuristics may be in place to facilitate its robust, efficient, and dynamic use.

This modelling perspective serves several ends. It adds to our understanding about how humans perform Theory of Mind and why they perform it as they do. It can help inform the design of the myriad AI systems that need to model people in order to interact with, or simulate, them. Finally, it may provide guidance in how AI systems can improve the robustness and efficiency of their own ToM reasoning. The following section outlines the structure of this thesis.

1.5 Thesis structure

Most of previous ToM research has studied either a only some elements of the computation of an inference (i.e. what brain areas are involved and how a mental model may be generated from existing knowledge), or the use of mental models in practice (i.e. the impact of mental state representations on the prediction of and interaction with others' actions). This thesis expands

the view on the underlying elements and aims to solidify a more detailed understanding of the processes and mechanisms that make up ToM. Specifically, it explores the relevant mechanisms from a perspective of cognitive costs: it identifies more concrete inferential patterns, heuristics that contribute to reduced cognitive costs, and relevant opportunity costs.

Chapter 2-4 of this thesis document the studies carried out as part of this work. These studies explore the processes and costs involved in ToM to form a basis that can inform more human-like computational modelling of ToM in the future. Chapter 2 focusses on the inferential patterns making up mental state representations, exploring the heuristics at the inference stage of mental model manifestation. Chapter 3 investigates stereotypes as an example of ToM heuristics and considers a perspective of cognitive costs in more depth. The chapter studies the factors contributing to what a mental model will contain, and the role of existing beliefs in shaping mental models and vice versa. Chapter 4 then establishes a wider angle on the factors contributing to robust, efficient, and dynamic ToM in social interactions. It studies how different sources of information shape the content of mental models, including the selective attention to a stimulus in the first place. The chapter expands the perspective on ToM to the aspect of its effect on behaviour. It also highlights individual differences and situational demands as a way to increase available cognitive strategies that drive efficient ToM. The following questions are more specific research questions that guide the individual chapters:

What cognitive dynamics and heuristics characterise efficient yet robust ToM? Sub-questions per chapter:

- Chapter 2: Heuristics in mental model manifestation
  - What cognitive dynamics and heuristics characterise efficient mental model manifestation?
  - What inferential patterns guide the manifestation of mental models?
- Chapter 3: stereotypes as heuristics in ToM
  - What is the role of stereotypes in shaping efficient ToM?
  - What factors affect the use of stereotypes in ToM?
  - What circumstances prompt mental model change?
- Chapter 4: ToM heuristics in interactions
  - Humans are complex and there are plenty of possible inferences to be made about others' mental states. What cognitive dynamics and heuristics affect the determination of specific inferences?
  - What cognitive dynamics and heuristics characterise ToM variety?
  - What cognitive dynamics and heuristics contribute to maximising ToM efficiency in social interactions?

This thesis has a particular focus on qualitative data as a way to illuminate ToM processes and mechanisms more than ToM outcomes. This qualitative angle highly contributed to the identification of common overarching patterns that characterise more nuanced and specialised ToM use. All findings are examined from a perspective of bounded rationality with the objective to minimise cognitive costs. The different chapters establish various processes, their costs, and trade-offs that are found relevant in shaping ToM manifestation and heuristics. The individual chapter discussions are followed by an overall discussion in chapter 5, including conceptual model propositions and future directions.

Questions

Blue boxes contain questions.

Insights

Green boxes contain insights.

This thesis contributes to the body of ToM literature by examining which heuristics may

characterise various elements involved in human ToM processing. Considering such heuristics may shape the conceptualisation of ToM overall and explanations for previously identified shortcomings of human mindreading abilities. Furthermore, this thesis generates new angles and hypotheses on ToM which may be tested in future approaches to modelling ToM.

Chapter 2

Heuristics in Mental Model Manifestation

2.1 Ad Hoc Theory of Mind

This chapter explores the structure underlying the dynamic nature of ToM. It demonstrates and argues for an ad hoc organisation of mental models.

2.1.1 Introduction

Mental models

Mental models are understood as a person's internal representations of the world. This includes knowledge about a system's characteristics, structure, or actions, to describe, predict, or explain its behaviour (Jonker et al., 2010; Rouse and Morris, 1986). The present use of the term describes mental models of other people.

Previous research highlights the nature of mental models as messy and incomplete (Norman, 1983). Rouse and Morris (1986) point to the role of generalisations informing the content of mental models and discuss differences between mental models and knowledge per se. They argue that, as opposed to assumptions in previous work, mental models are far from perfect, which needs to be considered in studying them. Assuming very analytical cognitive approaches to the generation of mental models is therefore not consistent with the complexity evident in both the brain and the social world.

Psychological perspectives

Mental models are crucial for social interactions. They help humans make sense of others' actions so that they can respond appropriately and predict their behaviour. Examples of the importance of functioning mental models come from research of what happens when they break down. For example, many autistic individuals have been found to struggle with the generation

of mental models of others', and commonly experience difficulties in social interactions (Baron-Cohen, 2000; Leppanen et al., 2018).

Research suggests that instead of constant elaborate processing, human brains make use of abstract and more general knowledge and theories about the world to guide the way in which new information and problems are interpreted and dealt with (Tenenbaum et al., 2011). The ability to switch between assimilation, i.e. integrating new information into existing knowledge structures, and accommodation, i.e. forming new knowledge structures when the existing ones are not sufficient, (Piaget, 1976) allows humans to learn from very sparse data and generalise to new situations. Caligiore et al. (2014) suggest that an emphasis on accommodation and an underuse of assimilation compared to neurotypical individuals may be a cognitive pattern in autistic people. They explain that the ability to generalise to different but similar stimuli and situations is a powerful tool to process the world effectively with minimal use of resources. This resonates with the common difficulty in autistic people to predict mental states or hidden meaning, such as non-verbal communication, sarcasm, or implicit information (Baron-Cohen, 2000).

Mental models are fundamental to effective teamwork (Mathieu et al., 2000). Similarity of mental models across the members of a team has been associated with better team performance (Bolstad and Endsley, 1999; Lim and Klein, 2006; Jonker et al., 2010). There is increasing interest in the study and use of mental models in AI research communities (Albrecht and Stone, 2018). The application of mental models in artificial agents has been shifted to the core of a lot of contemporary AI development.

Modelling perspectives

Many techniques used in contemporary AI to model others' mental states are very different to human cognitive processes. Models can differ in the methods they are based on, but all are characterised by performance very targeted to the task at hand (Albrecht and Stone, 2018). With rich data sets as basis and elaborate searching across different possible alternatives, these models can achieve high accuracy. However, the processes involved are costly and restrict the model's flexibility. The elaborate search algorithms and inferences involved in many computational models come with exponential increase of required processing as the reasoning process becomes more sophisticated and complex. Humans, in contrast, are proficient at learning general principles from very small samples of experiences and applying them to novel situations (Tenenbaum et al., 2011). From a modelling perspective, each cognitive process involved in ToM reasoning can be considered as a computation that has a cognitive cost attached to it. With a natural limit of resources and cognitive energy available, humans use patterns of heuristics to navigate the world effectively without spending too much energy on each one problem or task (Kahneman, 2011).

The efficiency characterising human brains is an attractive attribute to AI researchers, par-

ticularly for modelling processes as complex as others' mental states, potentially branching out to an endless number of possibilities. Striving towards AI that resembles humans has gained considerable interest (Lake et al., 2017) due to their accuracy, flexibility, adaptability, and efficiency when faced with novel situations and interacting with each other. The modelling of human mental models can also greatly inform the understanding of human cognition in general. In contrast to conceptual or computational models of mental models, human mental models are messy, fragmented, and often re-arranged quickly and reactively (Barsalou, 1983; Rouse and Morris, 1986). As Pynadath and Marsella (2007) show, mental models can be modelled and understood as minimal and highly targeted, specifically adapted, minimized, to serve the goals of the human modeling another.

Modelling mental models requires a high level of specificity in the understanding of what a mental model is, what it contains, and how it is formed. There is still a lot of work to be done towards the goal of modelling of human-like mental models. Specifically, this work is concerned with the mechanisms allowing for a balance of detailed yet efficiently generated mental models.

2.1.2 Rationale

The aim of this study is to improve the understanding of human mental model generation towards the longer-term goal of modelling the process. Specifically, this study targets the following two research questions:

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

- What cognitive dynamics and heuristics characterise efficient mental model manifestation?
- What inferential patterns guide the manifestation of mental models?

2.1.3 Methods

Participants

This study was approved by the ethics board of the Psychology department at the University of Glasgow. With the online platform Prolific 40 participants were recruited for this study. They were from the United Kingdom with at least 10 previously completed studies on Prolific and a 95% approval rate on Prolific.

Design

This study is of exploratory nature, asking participants about the mental states of 6 referents: *your mother*, *the ruling party*, *children in general*, *your best friend*, *a bus driver (in general*,

and *the police*, which were presented in a random order. These target referents were chosen to include many different relationships and capture differences between different types of mental models. Two dependent variables were measured, both in a qualitative format of open questions. The first of these two dependent variables consisted of examples of mental states in another person. The second variable was an explanation for why the respective mental state was held by that person.

Thornton et al. (2020) described mental states as the underlying drivers of the social world. In line with the discussion by Simons (1992), the more general term was broken down into more specific elements: *beliefs*, *goals*, *intentions*, *habits*, *preferences*, and *opinions*. Participants were asked to give examples of each mental state for each person of reference. They were then asked to explain why they thought the referent has this mental state.

Materials

The questionnaires were administered online, on the survey platform Qualtrics. Questions about others' mental states were asked in the following format:

> Give me a few examples of [your mother]'s beliefs.
>
> 1. Belief:_____.
>   Why do you think [your mother] has this belief?
> 2. Belief:_____.
>   Why do you think [your mother] has this belief?
> 3. Belief:_____.
>   Why do you think [your mother] has this belief?

Procedure

Participants chose on Prolific that they wanted to participate in this study and were re-directed to Qualtrics, where they gave informed consent and completed the questionnaire. They then received a full debrief and were re-directed back to Prolific where all participants received their payment when data collection was completed.

2.1.4 Results

Quantitative Results

Figure 2.1 shows counts of different categories how mental state attributions were explained.

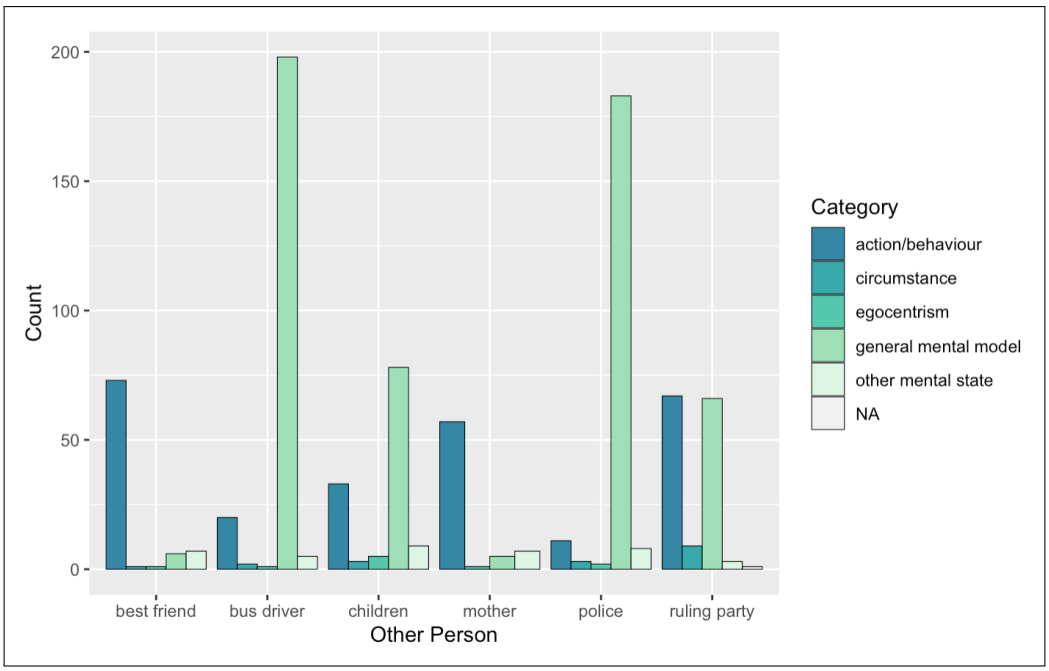


Figure 2.1: Counts of explanations for mental model inferences.

Of a total 865 attributions of others' mental states, 30% were explained with the other's actions or behaviour, 2% with external circumstances, 1% with egocentric explanations, 62% with general categorical knowledge, and 5% with another mental state attributed to the person. Figure 2.1 indicates that attributions based on actions were more common when people were known personally and explanations based on general conceptual knowledge were dominant when the target was a prototypical member of a larger social group.

Qualitative Results

Qualitative responses were analysed with focus on:

A) The efficient manifestation of mental models

B) Inferential patterns guiding the establishment of mental models

The 6 steps of the AMEE Guide No. 131 (Kiger and Varpio, 2020) were followed to achieve a thematic analysis of the data. Thereby, the researcher 1) familiarised themselves with the data, 2) generated initial codes, and 3) searched for overall themes. The themes were 4) reviewed and 5), defined and named, before all results were 6) written up in an overall report. In the following, the four themes that were found, and their respective sub-themes, are presented. Subsequently, the results are discussed and it is discussed how they may inform a conceptual model.

Theme 1: ***Model structure***

i. There are causal relationships between different mental states. Beliefs, opinions, preferences, goals, intentions, and habits can be inferred from behaviour.
[habit of best friend] ”Learning” – “Enrolled to a lot of courses”
[belief of best friend] “I believe in Green issues” – “Because she campaigns for this”
[intention of mother] “to be a good wife” – “because she is and this is important to her”
[opinion of mother] “Tattoos are ugly”- “She doesn’t want them herself”
[preference of best friend] “Blonde women” – “All his girlfriends were blonde”

ii. Inferences are often based on other inferences.
[intention of children, in general] “To save money” – “Adults tell them they can use the money to buy what they want and they want that independence”

The data suggests that very different types of mental states can be inferred from behaviour and that mental states can also be inferred from other, already inferred, mental states. One participant, for example, inferred the intention of their mother to be “to be a good wife”, “because she is and this is important to her”. To generate this prediction, both her behaviour and her values are referred to. The data suggests that different elements are used flexibly to make sense of others’ mental states and cognitive reasons for their behaviour. For example, sometimes preferences can be used to infer intentions, other times values can predict intentions. Depending on the information available, participants used it to express their thoughts about others’ mental states.

When people reason about one or several of these mental states, not all other mental states are reasoned about. Rather, the results at hand suggest that the activation of mental state representations is highly selective and minimal. Based on the results, the hypothesis is posed that human ToM includes the maintenance of coherence among different mental states. Mental states are inferred in such a way that is in harmony with other inferences and evaluations.

Theme 2: ***Filling gaps***

i. Mental models are based on sampled observations, which are generalised.
[preference of a bus driver, in general] “Drive during the day” – “Evening busses may have some difficult passengers to deal with”

ii. When no observations of behaviour are available, mental states can be inferred from a person’s role rather than directly from the individual.
[belief of a bus driver, in general] “public can be rude” – “sure they see it daily”
If only little information is available about a person, that information has more weight in influencing predictions about them.

[habit of a bus driver, in general] "Walks a lot" – "to [counteract] all the sitting down at work for their shifts"

iii. Even fictional information is used to predict others' mental states.
[preference of a bus driver, in general] "prefers being at home" – "he misses his family"

iv. When mental states are not known, there is the general assumption that people have the beliefs that are expected of them.
[opinion of children, in general] "They can do what they want" – "Because that's just children"

This theme suggests that the brain fills in missing information necessary for reasoning about others. Different strategies to achieve this were observed, such as generalising from similar situations or a person's social role. Interestingly, even fictional information was used to fill knowledge gaps. The proposition is put forward, that a considerable proportion of ToM consists of replacing knowledge gaps with available information that fits into them, even when it is not first-hand information or adopted from a different context.

Theme 3: ***People in general***

i. There is the general assumption that people seek out what they prefer and avoid effort or dislike.
[preference of police] "go after easy targets" – "quick wins and boost appearance of solving crime"

ii. There is the general assumption that behaviour is rooted in an underlying set of values and motivations.
[habit of mother] "Recycling and reducing" – "Every opportunity she looks to reduce water to help the planet"

Rather than fixed structural patterns of others' mental states (Theme 1), it appears that there are general assumptions about people's behavioural tendencies which help guide the understanding of others' actions. Specifically, regardless of the person, participants seem to assume that others seek out what they prefer and avoid what they dislike or what takes effort. There also seems to be a clear understanding that people generally act based on an underlying set of values and motivations. As part of a general mental model of a human, these assumptions can guide the processing of others' actions and reduce the resources required to make sense of social situations.

Theme 4: ***Ingroup-outgroup relationships***

i. Personal approval or disapproval affects the specificity of inferred intentions, leading to a more black-and-white or a more detailed perspective.

[belief of ruling party] "make the rich richer" – "BECAUSE THEY LOOK AFTER THEIR OWN"
[habit of ruling party] "efficiency" – "lowest unemployment rates in decades"

ii. There is the general assumption that a person who is similar or different in one way will be similar or different in other ways.
[goal of children, in general] "Stealing candy from the fridge" – "I had this goal as a child"
[intention of ruling party] "to have fun" – "Because they are elitist and can afford to have fun and break rules"

Finally, in-group or out-group membership appears to affect ToM reasoning. People who are similar in one way are predicted to be similar in other ways as well, whereas people who are different in one way are predicted to also be different in other circumstances. This provides an additional heuristic to reason about others and their mental states. If one can conclude that differences can be generalised to other situations or mental states, unless there is clear contrary information available, this reduces the resources required to infer a mental state. In the following, results will be discussed and insights informing the development of a conceptual model will be integrated with already existing literature.

2.1.5 Discussion

Psychological Perspectives

The first theme is concerned with the structure of mental models of others. It is a commonly accepted assumption that mental states, such as beliefs, opinions, preferences, goals, or intentions, are inferred from behaviour by means of backwards inference (e.g. Baker et al., 2011). The findings of this study suggest that different mental models are highly interconnected and dynamic. It appears that various types of information about the other person are recycled and generalised to different contexts by following the principle that mental states are overall consistent. By establishing overall coherence across the representations of different mental states, the accuracy of individual instances of ToM may be compromised, which has been observed in various studies, e.g. Keysar et al. (2003). On the other hand, it suggests a robust way to very reactively make sense of others' mental states and cognitive reasons for their behaviour. This insight is recommended to be tested in future confirmatory research with more stringent and controlled measures than the qualitative and exploratory approach at hand.

The harmony between different mental states that this suggests to be part of ToM is consistent with Thagard's (2002) accounts of coherence and Festinger's (1957) theory on cognitive dissonance, which are supported by robust bodies of evidence. An important implication for ToM reasoning is that representations of mental states may not be inferred from behaviour independently, but appear to be embedded in a larger network of knowledge about others. Most

interestingly, even when the other person is a stranger, the present results suggest that predicted mental states can be incredibly specific. This finding is novel to the research of ToM and will be explored more in future studies.

The second theme of this study taps into the brain's ability to deal with missing information. When a visual stimulus is incomplete, the human brain will draw on relevant knowledge to create a complete and consistent experience (Clark, 2013, 2015). For example, each human eye has a blind spot with no photo receptors where the optic nerve meets the eye. However, this lack of visual information is made up for by top-down influences from the brain, regarding what is most likely to be perceived at this point on the visual field (Wandell, 1995). Based on the present results, it is concluded that the ToM mechanism operates in a similar way, using already existing, higher-order, and more abstract knowledge to inform lower-level processes and ensure complete representations even when information is missing. Barsalou (2003) discusses simulation and abstraction as the result of repeated experience of different instances of a category. The influence of more abstract, general knowledge can then constrain new interpretations and learning (Tenenbaum et al., 2006). Interestingly, the results at hand suggest that the attributes to members of a general category are not general, but incredibly specific, even though there is only much more sparse and abstract information available. It can be argued be that ToM reasoning is not necessarily based on set mental models. Instead, it can be carried out in an ad hoc fashion, generated dynamically and reactively, depending on the situation.

Theme three offers another argument for less set and more ad hoc mental models. The theme suggests certain policies for human mental states in general. Policies are general rules that can be applied to guide the generation of inferences. Previous studies have connected this principle with ToM research to understand the mechanisms underlying mental models (Pöppel, 2023). Specifically, the findings suggest that participants use the general assumptions that others seek out what they prefer and avoid effort or dislike. Moreover, they reason on the basis that all behaviour is rooted in underlying values and motivations. These rules or policies may guide the process of filling gaps when less information is available, for example when a person is less familiar. Rather than waiting for a verbal statement of a preference, for example, participants can use other behaviours to infer the preference by assuming that their behaviour reflects a preference. The preference itself does not have to be stored in memory, it can be generated reactively when it is needed. Similarly, it can be argued that when reasoning about another person, not a complete mental model of that person is generated or "opened". Instead, only the relevant components are retrieved on an ad hoc basis.

The final theme is concerned with in-group and out-group membership to inform the generation of mental state predictions. Research on in-group and out-group dynamics is typically focussed on behaviour (Kurzban and Neuberg, 2015). The results draw attention to the potential impacts of group membership on the perception and representation of others in the first place. For example, members of an in-group appear to be associated with much more sophisticated and

detailed mental representations, whereas members of an out-group appear to be processed in a more black-and-white manner. Of course, it may be that they are in the in-group or out-group because of what is known about them. Hilton and Von Hippel (1996) however, highlight the potential of abstraction in representation hierarchies to enhance edges and reduce noise among different categories, which is what is observed in this data. Abstraction mechanisms, such as in-group and out-group distinctions, may be ways to organise more general mental models and reduce the extent to which they need to be sophisticated and detailed. Here it is concluded that these representations are then drawn upon on an ad hoc basis to reason about social situations reactively and dynamically.

Modelling Perspectives

The data was collected and analysed from a psychological perspective. Here a more concrete and modelling perspective is taken to interpret the findings from a standpoint of aiming to model ToM. This section extracts the components of the results that can aid conceptual modelling of the phenomena at hand. The insights of this chapter are primarily about the ToM inferences themselves.

Firstly, Table 2.1 shows inferential patterns that are proposed based on the results above, with inferences highlighted in green and given explanations for those inferences in blue.

Behaviour x → outcome x → goal x
Behaviour x → outcome x → plan x → intention x
Behaviour x → x is true → belief x
Regular behaviour x → habit x
Behaviour x → subjective evaluation x → opinion x
Behaviour x → choice of x over y → preference for x

Table 2.1: Proposed inferential patterns in ToM, including Explanations and Inferences.

The results of this study suggest many different inferences drawn from representations of behaviour. However, there are patterns of certain characteristics being mapped to different underlying mental states. For example, intentions were inferred from behaviour by reasoning about the achieved outcome and the underlying plan, whereas habits were inferred by reasoning about what behaviours were regular. Table 3.1 suggest different mappings between observations of behaviour and representations of underlying mental states. This may provide more specific

grounds for finding patterns in ToM inferences. These proposed inferential patterns may be tested in future research.

Secondly, various elements of the results presented above suggest an ad hoc approach to ToM inferences. This encapsulates the phenomenon of putting together a mental representation of another person in the moment from relevant other existing mental representations, rather than retrieving or creating an existing unique mental model of that individual. In modelling terms, not all elements preceding the mental model in Figure 3.1 may be executed all the time. Rather, it is possible to only retrieve existing beliefs about the social group an individual belongs to or the general situation they are in, and not infer a person's mental states from their own behaviour. That way, the mental models stored in memory do not have to be incredibly specific and available in a large number but can be generalised to different people.

A final take on inference generation from a modelling angle are the following principles extracted from all qualitative themes collectively. The results suggest overall heuristics that constrain the possibility of mental states underlying a behaviour:

- People seek out what they like and avoid what they dislike
- Behaviour is rooted in an underlying set of values and motivations
- A person who is similar or different in one way is similar or different in other ways
- Mental states align with each other and with the person's behaviour

These heuristics are suggested to constrain the space of inferences that can be made about another person's mental states. In terms of the different elements of ToM investigated here, the possible heuristics suggested above fit into the element of mental state representation (highlighted in Figure 2.2).

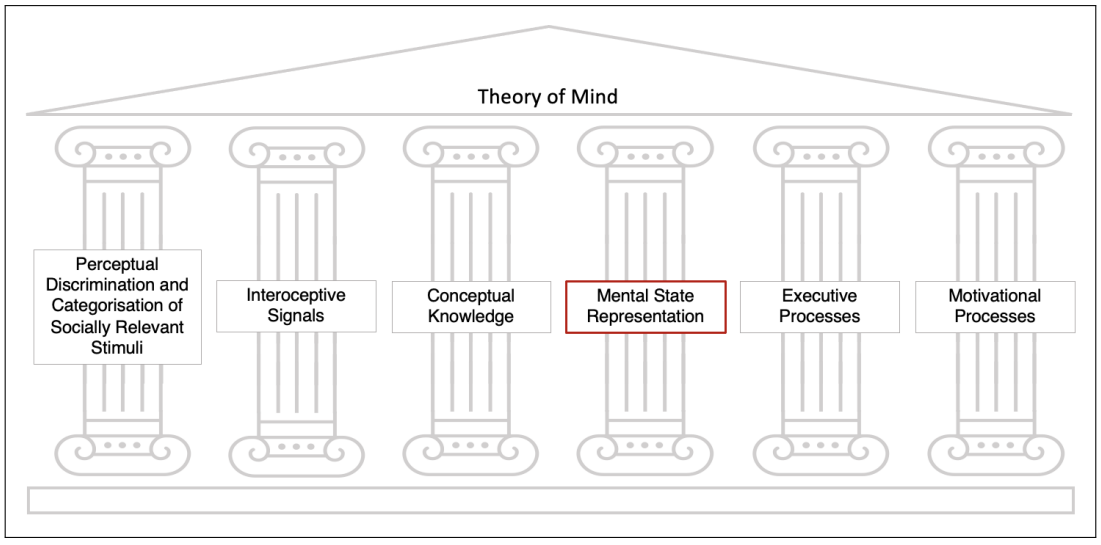


Figure 2.2: ToM elements considered in this thesis, based on Schaafsma et al. (2015).

2.2 Insights

Recall the research questions for this chapter.

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

- What cognitive dynamics and heuristics characterise efficient mental model manifestation?
- What inferential patterns guide the manifestation of mental models?

The following summarises the heuristics proposed based on the results of this study:

Heuristics in Mental Model Manifestation

- Mental models can be generated on an ad hoc basis
- Different mental states can be inferred from different features of representations of another person's behaviour
  - Behaviour x → outcome x → goal x
  - Behaviour x → outcome x → plan x → intention x
  - Behaviour x → x is true → belief x
  - Regular behaviour x → habit x
  - Behaviour x → subjective evaluation x → opinion x
  - Behaviour x → choice of x over y → preference for x
- There are general principles that guide the determination of mental state inferences
  - People seek out what they like and avoid what they dislike
  - Behaviour is rooted in an underlying set of values and motivations
  - A person who is similar or different in one way is similar or different in other ways
  - Mental states align with each other and with the person's behaviour

As mentioned in the introduction, this work considers cognitive energy as an indicator of cognitive costs. The qualitative results above are possible explanations of how cognitive energy expenditure may be reduced at the stage of the mental model formation. To form a mental model, 1) relevant information needs to be considered to 2) then make an inference about what is most likely to be the mental state. The insights above suggests that cognitive costs may be reduced at the stage of mental model manifestation in various ways.

Firstly, in the broader context of ToM mechanisms and processes (Figure 3.1), the results of this study suggest that mental models are not necessarily formed in an encapsulated fashion and stored in memory separately for each individual a person encounters. Rather, it is hypothesised that different elements from more general mental models can be combined on demand to form ad hoc mental state representations. This demand-driven approach suggests a larger focus on information search or retrieval from different relevant belief structures and less cognitive energy spent on the storage and maintenance of complete and rich mental models. It also indicates an influence from multiple sources of mental model content which can be re-used and combined. Here it is proposed that there is a simple more general framework to which details are pulled in when required and/or available. This phenomenon may have several benefits. It may reduce the cognitive energy that would otherwise be required to acquire and remember information needed to maintain the model. It also allows for flexible and dynamic generation of mental models from applicable sources. This dynamic approach may therefore be more robust when novel situations occur. Secondly, the chapter suggests concrete links between different features of observed behaviour suggesting certain underlying mental states. This mapping will be useful in modelling ToM. Finally, inference generation may be guided and facilitated by general heuristics such as the assumption that people are consistent, enhancement of contrasts and edges, and the principle that others act based on goals and preferences. This work proposes that following these heuristics rather than having to carefully establish patterns for each individual a person will meet, can greatly reduce the cognitive energy spent on ToM.

These results are of qualitative nature and indicate new ways of considering the manifestation of mental models in a cost-efficient manner. Future research may test these hypotheses quantitatively to confirm whether the above considerations are fruitful for the conceptualisation of ToM.

Chapter 3

Stereotypes as Heuristics in Theory of Mind

This chapter investigates the processes underlying the generation of mental models. It explores the factors contributing to mental model content as well as mental model change, and considers the role of existing beliefs, including stereotypes, in this process. It takes an approach to understand relevant mechanisms from a perspective of cognitive costs.

3.1 The Human Library

3.1.1 Human Books

The Human Library is an organisation specialised on breaking down stereotypes and biases (Giesler, 2022). Their events include opportunities of speaking to “human books”, people who belong to stigmatised social groups and experience bias and prejudice. The "readers" can speak to a person subject to stereotypes and stigma to explore and break down their own prejudices. The opportunity to have a conversation with these books has been found an effective way for participants to learn more about a book’s perspectives and challenge existing biases and stereotypes (Groyecka et al., 2019; Watson, 2015).

Stereotypes are generally defined as generalisations about a person’s traits based on their membership of a social group (Allport et al., 1954). Inferring beliefs about a person from their group membership is a powerful source of information in ToM (Westra, 2019). The overlap of stereotypes and ToM has been considered, however, it has not been studied extensively (Mulvey et al., 2016; Westra, 2019). In the context of efficient ToM, stereotypes may serve as more generic sources of information to use when a mental state inference is needed but first-hand observations of the individual are limited. Madva and Brownstein (2018) suggest that there are complex but tight relations between stereotypes, concepts, beliefs, and affect. Interestingly, Rusch et al. (2020) propose that the use of ToM increases when social situations are more un-

certain or interactive, which highlights the potential of illuminating both ToM and stereotypes in the domain at hand.

The Human Library interventions create a very interesting environment of exploring and challenging stereotypes, which at the same time is an opportunity to closely study the role of stereotypes in ToM. This work is a collaborative effort to explore stereotypes as potentially cost-saving heuristics in ToM.

3.1.2 Costly Cognition

Increasingly, computational models of virtual agents and/or human behaviour aim to include mental models (Bansal et al., 2019; Pynadath and Marsella, 2005; Gmytrasiewicz and Doshi, 2004), particularly considering the ever-increasing rise of HCI (human computer interaction). Modelling humans and other agents allows a virtual agent to predict others' actions and respond to individual needs (Jameson, 1996).

From a modelling perspective it is useful to detail what makes ToM reasoning costly and how processing costs can be reduced to make practical implementations possible and tractable. At its core, ToM is the representation of another's mental states, which involves establishing and maintaining such representations. It requires integrating new observations and contextualising another's behaviour, potentially sequences of behaviour over time. The ways or mappings by which inferences are drawn from observable behaviour or events also need to be stored and maintained as representations, which requires cognitive resources. Moreover, integrating new observations with existing representations involves the adjustment of these representations as and when required. Furthermore, the representation of another's mental states is characterised by ambiguity, uncertainty, and variety. Representing large numbers of possible mental states that could apply to an individual in a given situation would be very costly for a person to process and conflicts with the speed and ease with which humans use ToM. All these parameters suggest that there are mechanisms that constrain the ways in which ToM is performed. This work is placed in the broader context of understanding the cost-saving mechanisms that characterise human ToM as a potential way to achieve cost-efficient modelling of ToM.

It has been suggested that ToM in artificial agents can be made more efficient by using minimal models, restricted by the observing agent's goals and potentially similar to stereotypes (Pynadath and Marsella, 2007). There are likely other heuristics contributing to the sophisticated yet not cost-efficient use of human ToM, which have been pointed to in theory (Gallagher and Fiebich, 2019) but have not been specified mechanistically to the extent that they could be modelled computationally. Humans have no direct access to others' mental states. Regardless, they draw inferences to understand and predict others, and social interaction is often fundamentally characterised by these assumptions (Rusch et al., 2020). Therefore, many mental models rely on inferences and assumptions to a considerable extent (Koster-Hale and Saxe, 2013). Finally, the typical manifestation of meta-representations is still to be studied. There is robust evidence

that the representation of the physical world is hierarchically structured and characterised by associations (Zweig and Weinshall, 2007). It is yet to be determined whether this also applies to the social world.

ToM has often been approached in a "black box" manner: Inputs and outputs have been studied extensively but underlying processes and mechanisms are yet to be explored and clarified (Schaafsma et al., 2015). This study aims to investigate the structure and processes associated with mental model representations more closely. Specifically, it considers a scenario of stereotypes and prejudice, reflecting the complexity of the social world, to illuminate and document what happens when a mental model is altered. In the context of the bounded rationality principle (Simon, 1956; Lieder and Griffiths, 2020), in which the availability of cognitive resources fundamentally contributes to the degree to which a decision or action is optimal, stereotypes are considered as a type of heuristic to reduce the expenditure of cognitive effort while maintaining a sufficiently high level of accuracy in understanding, representing, and predicting the social world.

The subsequent sections document how participants' mental models are affected by attending a Human Library session. Insights from before and after participants attended a session are compared. This is discussed with regards to the role of stereotypes as an entry point to understanding the processes underlying social inferences and the factors shaping mental models. Particular attention is be drawn to the social and computational costs of ToM, what cognitive patterns may reduce these costs, and how understanding these patterns can inform the modelling of social cognition.

3.1.3 Rationale

This work reports on a collaboration with the Human Library, in which the role of stereotypes in shaping mental models as a cognitive shortcut was explored. The research was conducted qualitatively to investigate the issues described above in detail to inform a conceptual model of the issues at hand. The study was guided by the following research questions:

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

- What is the role of stereotypes in shaping efficient ToM?
- What factors affect the use of stereotypes in ToM?
- What circumstances prompt mental model change?

3.1.4 Methods

Participants

A participation invite to this study was sent out with help from the Human Library Organisation to all the people who had signed up to attend their session on the following weekend. Response emails were answered in the order they were received in, to arrange one interview before and one after the Human Library session with each of a total of 12 participants. 11 of these scheduled interviews were successfully carried out and recorded. Participants varied in their Nationalities and ages.

Design

The nature of this study was qualitative and included semi-structured interviews.

Materials

In preparation, the researchers received a list of the human books which would likely be available to participants during the session. This included *Gay*, *Obesity surgery*, *Vitiligo (the skin condition)*, *Dominatrix*, *Intersex*, *Bi-Racial first generation American*, *Genocide survivor against Tutsi*, *Black American*, *Lesbian*, *Bereaved by suicide*, *Survivor of domestic abuse*, *Disabled kid*, *Transplant recipient*, *Immigrant*, *Pagan*, *Cancer*, and *Epilepsy*. Not all but most book titles were available in the sessions. The Human Library staff determined which participant would speak to which books.

Interviews with the participants were guided by the overall research questions stated above. Specific interview questions included:

These questions were chosen to explore the role of stereotypes in the formation and change of mental models. Based on Qu and Dumay (2011), the questions were used to elicit the participants' own stories and experiences while still keeping the focus on mental models. The questions in interview 1, before the readers attended the Human Library session, were focussed on establishing what stereotypes were being held before participants spoke to the human books. In interview 2, these representations were already affected by the interactions with the human books. The questions in the second interview were therefore focussed on the experience of developing new or different representations due to their Human Library participation. Overall, the collective questions from both interviews were targeted to capture an overall narrative of how mental models were affected by the readers' participation in the human library session. The interviews were kept open, using prompts such as "Can you elaborate on this?", "Why do you think that?", "How did you pick up on this information?", or "How did that make you feel?" to encourage detail and reflection. Otherwise, participants were encouraged to speak freely about their expectations, perceptions, and experiences.

Interview 1	Interview 2
• What immediately comes to mind if I say [book title]? • What do you think their thoughts and feelings might be? • What does it mean about them? • What do you think what assumptions might be out there about them? • What do you think they would think about you? • What question would you ask them?	• What did you expect going into the conversation? • Did anything surprise you? • What was your main take-away from this conversation? • If you were to meet another [book title], would you feel or think any differently about them?

Table 3.1: Interview Questions.

Procedure

Before their first interview, each participant received an information sheet and consent form. The consent form was signed and returned before the interviews took place. All interviews were carried out online. Participants could decide if they wanted to turn their camera on or off. With their permission, the videos were recorded. The first interview was scheduled before the participant took part in the Human Library session, in which each participant spoke to two different books. The session itself was organised by the Human Library Organisation who also organised the allocations of all participants to a selection of human books. The second interview was scheduled for after the session. Participants then received a full debrief and a payment of £20 for the two interviews to their bank account. Two participants wished for their payment to be donated to the Human Library Organisation.

Interview recordings were transcribed and then destroyed, leaving only written transcripts. As agreed in advance, the transcripts were shared with the Human Library Organisation. All data was analysed with a thematic analysis, based on the procedures described by Kiger and Varpio (2020). This included the six steps of 1) data familiarisation, 2) initial code generation, 3) theme search, 4) theme review, 5) defining and naming of themes, and 6) reporting the results.

3.1.5 Results

The data was analysed with focus on: A) What determines the structure and content of a mental model? B) How can mental models be changed? C) Why do humans develop stereotypes and

what is their role in shaping efficient ToM? The AMEE Guide No. 131 (Kiger and Varpio, 2020) was followed step-by-step to achieve a thematic analysis of the data. In the following, three themes and their respective sub-themes are presented. It is noted whether quotes are from before (*) or after (**) a participant took part in the human library session.

Theme 1: ***Structure of mental models***

i. *Level of detail (black & white vs everybody is different)*
The present data suggests stark differences in the level of detail of mental models. Thereby, it appears that a mental model of an unfamiliar and/or stigmatised social group member is typically general, homogenous, and consistent. Group members are perceived to be similar in more ways than just their group membership.
"I would imagine that they have, have all, do struggle with food, and they therefore have a, have an eating, some sort of eating disorder some sort of binge disorder" *
"this is the assumptions that we make about people, again, is that we just see either the disability or we just see, you know, intersex, or we just see the tattoos. And we assume that everybody who's like that is exactly the same and they're not." **
Encouraging perspective-taking by means of open conversation, in contrast, seems to shift the focus from the generalisation to individual differences and the complexity of unique experiences.
"I wanted clarity. But [. . .] I was introduced to more perspectives" **
"what I took away from it is that a normal well-adjusted kid could also be a pagan" **
Participants indicated surprise and discomfort about clear boundaries being broken down. Maintaining simplicity appears less effortful and more intuitive than allowing cognitive dissonance by appreciating differences and complexity.
"the first impression will be like, oh, there's a, there's a real person behind the title, you know, so it will be like oh they're quite, quite a regular person, you know" **
"I kept myself to myself, [. . .] within my family, within the group of friends that I knew. And that was it, I was in this little bubble. Because that's what kept me confident" **

ii. *Categories and reference points*
Individuals described intuitive and automatic categorisation of new stimuli into existing memory structures and reference points as a way to interpret and understand the social world.
"I kind of automatically categorised them" **
"I suppose if I, if I see another intersex person then maybe yeah, well that will be my only reference point" **
Thereby, this existing structure was described as a fundamental component of interpreting social situations, such as predicting others' mental states or future actions.
"I guess, in order to understand, it's more of just having another thing to compare it with."

**

"I haven't met anyone [...] personally and I don't know what they going through, so I can't compare with, compare them with normal people" *

iii. *Egocentrism*

A common element of processing and interpreting others was that of egocentric bias. This describes the projection and generalisation of an individual's own knowledge or experience onto another person to understand or predict their actions.

"Why do you find it difficult to understand?" – "Because it wouldn't be anything for me" *

"Uhm, I think, I don't know, because I feel like that, I would assume everyone would feel like that" *

"well I know from my experience I would, I would be making assumptions that they're probably quite self-conscious. [...] nothing else that springs to mind except, I just know how I felt. When I had it." *

Many participants related another person's perspective to their own experience in order to accept it, as opposed to treating them without any expectations or pre-conceptions.

Overall, Theme 1 shows that high levels of detail in mental models are possible but not necessary for them to be applied and relied upon. Moreover, they are highly embedded in a person's conceptualisation of the social world, and influenced by their own, often unrelated, experience. It appears to be common for individuals to be very confident in their judgement even when very little data is available.

Theme 2: ***Reasons for judgement***

i. *Misinformation (from the media)*

In explaining where assumptions and biases about others came from, many participants referred to the media as a source of information. Responses show the extent of reliance on and confidence in second-hand (and commonly false) information as a way to learn about others.

"Documentaries and things like that, [...] it's probably the media creating the idea of this is what that type of person does or thinks" *

"I think it's, well, I suppose from the minute we're born we're conditioned with what we see, what your parents believe or whoever it is that you're brought up with, with the school that you go to, with the group of friends that you hang about with" **

ii. *Causal inferences from roles to traits*

Inferences were commonly drawn not from previously observed behaviours but solely from a person's role. Interestingly, the predictions about what a person with a specific

role or label would be like, were typically very detailed and specific.

"I don't think, when you say dominatrix, I don't think about someone who's gentle. Uhm I just instantly think about someone who's quite angry, aggressive, and enjoys inflicting pain. [...] I imagine them being single or perhaps attract a certain type of person. And perhaps maybe struggle in relationships a little bit." *

"[A disabled child is] probably the centre of their parents', attention, and they may well be a little bit, I don't know, a little bit spoiled because of that" *

iii. *Lack of information*

Many stereotypes and assumptions were talked about in relation to a lack of knowledge.

"I find it hard to find it difficult to understand, that is I don't understand it at all." *

Participants expressed uneasiness over being uninformed and having an open conversation about topics they were uncertain about. In contrast, there was a sense of positive surprise when such a conversation was carried out successfully.

"it's an uncomfortable feeling that I have. More so because I don't know enough about that. [...] And it causes me real anxiety. [...] The last thing I ever want to do is offend anyone." *

"Since I found it very easy to talk to him and ask my questions, I think I might be more open to people of colour also here in Germany. Because it was so easy" **

iv. *Fear of the unknown*

Beyond this, many participants described themselves or people they knew as fearful of novel and unconventional concepts or lifestyles.

"Could be a little bit [...] weird, you know [...]. Yeah, little bit frightening, because it's unknown, for some" **

"I would have been quite fearful probably, again maybe, [...] when we spoke about epilepsy, there's sort of the fear of the unknown" **

"my impression was before talking to him that a kid who stayed by himself, and did all this research on paganism and said he was a pagan was some kind of a weirdo freak, who might turn violent. That was actually what I thought. I thought this kind of kid who's [...] on the computer all the time. Watch out. That, you know, this is the kind of kid that I could see walking into a mall and shooting somebody" **

"I observed this in my friends and so on they are, they are feeling a little bit threatened [by] other ways of living that are not the same as [theirs] [...]. You know, people are harsh [...] when they feel that their way of living is in danger" **

Theme 3: ***Change of mental models***

i. *Challenging expectations*

Many participants reported that their participation in the Human Library session was very

different to what they had previously expected or imagined. Similarly, what they learned about the human books was often in contrast with their previous expectations and highlights the potential of the Human Library as an educative organisation.

"the conversation unfolded completely different to how I could imagine it" **

"I had never thought about that before. I had no idea that when you have a kidney transplant, they don't take out the old kidney. It just, that absolutely blew my mind" **

ii. *Shifting from generalised mental models to the complexity of individual experiences*

It appears that the learning accomplished by participating in Human Library sessions shifted many perceptions of social groups from more generalised conceptions to an understanding of individual differences. Many participants appreciated the complexity of individual experiences and reported that speaking to the human books contributed to this shift in perception and understanding.

"I think it's more complex than that. It's much more complex than I think I would initially have thought" **

"I actually think I felt more confused than when I went in, because we're all human beings, we're all unique and what might offend or upset one person might not necessarily upset or offend another person" **

"[Speaking to a gay person in the future,] maybe I'm trying not [to] worry about that. Maybe it's not, you know, it's not [...] the central point in human interaction if you're gay or not." **

"I feel like I know a little bit more about it now, but couldn't assume because everybody's still so different. So even, even somebody who's following a different religion that's a very strict religion and everybody's the same they're still an individual within that religion" **

"It's not really changed the perspective is more like widening space, you know [...] So, yes, but it gets more subtle, you get a more accurate approach. [...] it's a whole war, you know, a whole universe" **

iii. *Negative attributions change to positive attributions*

Overall, many attributions changed from negative to positive. For example, one participant changed their impressions from

"[a disabled kid is] probably the centre of their parents', attention, and they may well be a little bit, I don't know, a little bit spoiled because of that. [...] That's not to say they don't have huge struggles in their life, but, but probably their personality is a bit more entitled, because, I don't know" *

before the interview to

"because her parents didn't speak English that well, uhm, they didn't always understand what the doctors and the specialists were saying. And she, when she was explaining to them, she left out things that she didn't want them to know or that she was trying to protect

them from. So that was really different. The other way around, but you know, this, this young person, she must have been, you know, in her mid-teens [...] had the weight on her shoulders to protect her parents. [...] Uhm, just how brave and how... she was so optimistic" **

after the interview.

3.1.6 Discussion

Psychological Perspectives

The discovered themes are now individually discussed with regards to previous literature and considered from a perspective of cognitive costs. On one hand, this work aims to gain psychological insights. On the other hand, it aims to contribute to the body of AI research that aims to establish how human behaviour and human mental states can be modelled tractably and efficiently. Particular attention is drawn to the social and cognitive costs of ToM and how understanding these patterns can inform the development of conceptual and eventually computational models of social cognition. Implications and limitations are discussed.

Theme 1, Structure of mental models, highlights the fundamental role of previous knowledge structures in understanding and processing social stimuli. Already existing reference points appear to be necessary rather than simply beneficial for the assimilation of new stimuli. Furthermore, the data suggests that it is intuitive to think in simple terms and that basic mental models are commonly favoured over complex ones. Recognising the complexity that characterises individual human experiences is in direct contrast to stereotyping and generalisation. Stereotypes are a widely studied phenomena but there is limited research on its role as an entry point to understanding the mechanisms contributing efficient social inference generation. The present work starts bridging this gap in the literature. The finding that humans are biased towards simplicity and relatability in the generation and maintenance of mental models helps explain why it is so difficult to replace them, and to challenge existing stereotypes (Duehr and Bono, 2006; Hill and Augoustinos, 2001).

Moreover, egocentric perspectives appear to strongly influence ToM by means of relating others' experiences and behaviours to one's own. Importantly, in a situation when the own experience is inherently different to another's but used to predict or understand them, a person's conclusions can be very inaccurate. If an individual only has little or wrong information available about another person but the construction of mental models happens automatically as part of the process of predicting one's surroundings (Schneider et al., 2017; Freundlieb et al., 2015; Koster-Hale and Saxe, 2013), it is not surprising that poor representations can develop consistently over time.

Furthermore, unbiased interpretation of new social stimuli appears to require more energy than interpreting them by re-using already existing mental representations. The present findings

suggest that this can be achieved by means of egocentrism or stereotypes. This is in line with the "like me" hypothesis (Meltzoff, 2007) and its extension, the "like others" hypothesis (Bianco, 2022), which propose that observations and knowledge of one's own and others' actions are used as templates to understand future situations.

Theme 2, Reasons for judgement, highlights different sources of assumptions and stereotypes, and points to the risk of inaccuracy in ToM. Participants indicated misinformation or lack of information as contributors, but also false inferences and bias based on fear. The results at hand suggest that information from removed sources as opposed to primary sources is commonly taken at face value and typically not questioned or dismissed. Previous literature has suggested feelings of discomfort about uncertainty (Tanovic et al., 2018) and that this is linked to maintaining a sense of control (Mushtaq et al., 2011; Penrod, 2007; Edwards and Weary, 1998). Furthermore, stereotypes commonly have an element of fear associated with them (e.g. Low & Purwaningrum, 2020).

With a tendency to continuously predict the environment (Koster-Hale and Saxe, 2013), including social stimuli, it makes sense that a person would find a lack of knowledge uncomfortable and may prefer accepting (false) knowledge unquestioningly. This again can facilitate the development of inaccurate mental models and the social transmission of stereotypes. Interestingly, ToM has been suggested to be used more when social situations are more uncertain and interactive (Rusch et al., 2020). If a lack of knowledge increases amounts of ToM but also decreases the accuracy of predictions and interpretations, ToM naturally occurs in a space of little evidence and reliability.

Theme 3, Change of mental models, suggests attending Human Library sessions to be effective in changing participants' perspectives and accounts of members of stigmatised groups. Participants reported learning to appreciate complexity of individual experiences and that their models typically became more positive from learning more about a person. The findings indicate that this process typically involved challenging expectation and experiencing a degree of cognitive dissonance.

Human Library participation has previously been reported as an effective way to challenge stigma and reduce stereotypes (Groyecka et al., 2019; Watson, 2015). The findings at hand are in line with the contact hypothesis (Allport et al., 1954), which suggests that stereotypes and prejudice can be reduced by encouragement of interpersonal contact between groups (Pettigrew and Tropp, 2006). The present study highlights the Human Library as a promising approach to reduce stereotypes and provides more insight into why this might be the case. The results indicate that the generalisations that make up stereotypes were to an extent replaced with more detailed representations of the stigmatised social group, with the focus shifting to the appreciation of complex individual experience. This work documents a process of learning to establish more nuanced mental models and gathering more diverse information about a stigmatised topic.

However, regarding the overall generalisability of this study, it is worth noting that it does not

illuminate to what extent the observed learning is possible in one or several Human Library sessions, what previous background or attitudes may be necessary, and how learning may vary with individual differences. The people who attended the Human Library session and participated in this study all actively sought out the event to educate themselves and reported an already existing level of openness. Alongside building knowledge about a stigmatised topic and replacing generalised information with more detailed accounts, the learning achieved in Human Library sessions may reduce the extent of uncertainty and potential discomfort associated with it. This again highlights and supports the notion of a tight connection between uncertainty, stereotypes, and ToM in general (Madva and Brownstein, 2018). However, without following up with participants at a later point in time, it cannot be known if the reported changes in perception and evaluation are long-term or only temporary.

In summary, this study further illuminated human ToM by means of exploring mental model structure and change. Insight into stereotypes and mental model characteristics from before and after participants attended a Human Library session were reported, analysed, and discussed. Results illuminate the structure of mental models, highlighting the role of different levels of detail, categories, reference points, and egocentric bias. Findings also show different reasons for judgement of marginalised groups and illustrate cases of changing mental models in practice.

Modelling Perspectives

Insights in the structures and processes of reasoning about others are helpful for understanding ToM from an angle of psychological theory but also for driving modelling understandings of ToM. For example, to implement the discovered patterns in a concrete simulation, this model needs to be able to represent different levels of complexity in ToM, perform processes of generalisation, use previous information as reference points, and reflect the tendency to prioritise self-relating knowledge. The results and insights above can help specify underlying ToM processes further.

As specified in previous chapters, this work considers a perspective of costs and benefits to establish a more model-focussed angle on the issues at hand. With this standpoint, theme 1 suggests that humans may save cognitive costs by drawing upon already existing and generalised mappings rather than representing and processing mental models in detail. Stereotypes are a type of belief structure very commonly employed in ToM processes to reduce the amount of energy required to generate inferences about others' mental states. Using stereotypes has the benefit that new formation of mental models can be avoided. Stereotypes are also simpler to process than more complex and individualised representations of others. The use of stereotypes may therefore reduce the cognitive energy required to search for appropriate information and store a variety of nuanced and complex beliefs. On the other hand, this faster and more efficient process is likely to coincide with decreased accuracy of mental models. Future research may test this perspective of cognitive costs by modelling stereotype-based and comparing how much energy

it requires compared to non-stereotypical ToM.

Many approaches to the computational modelling of ToM are based on probabilistic inferences (Baker et al., 2011), accounting for uncertainty and missing pieces of information. It has been argued that the brain works like a "prediction machine" (Clark, 2013) and uses predictive coding mechanisms to accurately model the world. These approaches match the notion of understanding social stimuli with incomplete information found in the present results and highlights why probabilistic models may be a successful tool to implement ToM computationally.

The observations in the second theme can also be understood from a perspective of costs and benefits, specifically social utility. The results suggest social and situation-dependent costs of prediction errors, affecting the demands on the mental model. This may be conceptualised in a social reward function, shaped by an evolutionary development of ToM use (Heyes and Frith, 2014). The findings above indicate a general desire to predict and understand the environment, and to avoid uncertainty and ignorance. In some situations, therefore, a complete lack of knowledge may be more costly for the individual than sparse or inaccurate knowledge, even if this can lead to wrong conclusions. The results at hand also suggest that losing group membership is very costly, which is consistent with other literature on the relevance of social groups (Jetten et al., 2015; Mullin and Hogg, 1999). Maintaining beliefs and representations consistent with the in-group as a means of maintaining group membership, may therefore be prioritised over accurate or conflicting representation of others. The theme suggests that assuming a threat where it does not really exist is less costly than not perceiving a threat that poses real danger. This highlights the importance of uncertainty as an element affecting human reasoning patterns in social situations which is recommended to be encapsulated into a simulation of ToM.

Finally, theme 3 highlights that the process of breaking down stigma builds on the development of more complex and more costly mental models, which requires more energy than maintaining simple mental models and using egocentrism or stereotypes. The cognitive dissonance created when expectations are challenged and the experience of changing an existing beliefs are typically uncomfortable (Elliot and Devine, 1994) and cognitively costly. These processes are therefore likely considered worth the investment of cognitive resources for a beneficial reason. Many participants reported a shift from a negative to a positive attitude about the stigmatised group in focus. Maintaining positive, familiar, and non-threatening mental models as opposed to negative and uncertain representations of potentially threatening individuals, may be a benefit to outweigh the cost of experiencing cognitive dissonance and establishing more complex and costly mental models.

The explanation of cognitive costs as an explanation for the patterns observed has surfaced in many different ways in the discussion above. Therefore it is proposed that a model may benefit from elements of both cognitive and social costs and utility. If tested and successfully established, this may eventually also be an excellent research tool to simulate the learning observed in Human Library sessions and explore what conditions and factors facilitate the process

of breaking down stigma and stereotypes most successfully.

3.2 Insights

The Human Library study offers various new perspectives about the use of stereotypes and reference points, reasons for judgements, and issues around the change of existing and stored mental models.

Recall the questions for this chapter:

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

- What is the role of stereotypes in shaping efficient ToM?
- What factors affect the use of stereotypes in ToM?
- What circumstances prompt mental model change?

In summary, these may be answered in the following way:

Stereotypes as Heuristics in Theory of Mind

- Mental model content is tied to social costs (group membership) and preparedness (predicting environmental threats)
- Informing mental models with existing and generalised mappings is less cognitively costly
  - Beliefs can be re-used rather than having to be newly formed
  - Stereotype-based mental models are generalised and therefore simpler to process
- The demands to change existing belief structures and mental models need to outweigh the required cognitive effort
  - Inconsistency across mental models is uncomfortable
  - Cognitive restructuring requires cognitive energy
  - Breaking down stereotypes involves representation of complex individual experiences

In the context of the different ToM elements , these insights fit into the components of percepts, exisiting knowledge, and motivations (highlighted in Figure 3.1).

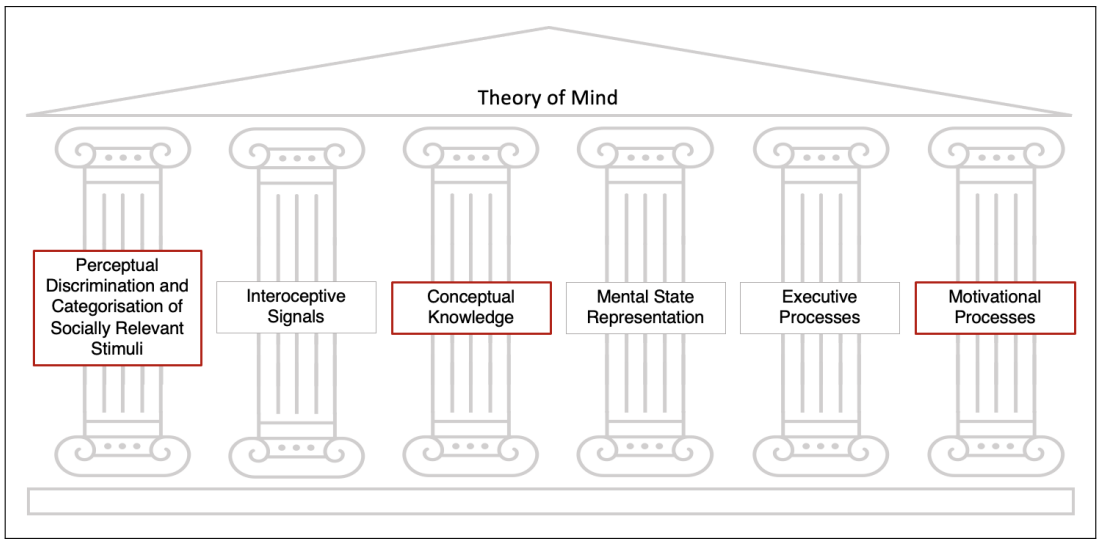


Figure 3.1: ToM elements considered in this thesis, based on Schaafsma et al. (2015).

This study particularly looked at the role of stereotypes as a heuristic to reduce cognitive costs and facilitate efficiency in the process of drawing upon existing belief structures to inform mental models. The results illuminate stereotypes as simpler and more generalised belief structures. It is proposed that they can speed up the process of forming representations of others' mental states by means of providing more general information that can inform mental models without the costs associated with individualised details about each and every other human a person will ever think about. It is also suggested that they reduce the need of complex re-structuring of existing beliefs, as they appear to be maintained at a simpler standard. On the flip side of this potentially energy saving mechanism, it appears that it can be quite difficult to change existing stereotypes.

Furthermore, the results of this study point to a potential benefit of including both personal and social utility in a model of ToM. The results reflect patterns of discomfort around uncertainty and social conflict. Generally, humans are motivated to maintain beliefs that are coherent and favour a person's in-group. This implies effects on the ways in which a mental model is pieced together and how the different elements that can shape a mental model may conflict with each other. This may look very differently for different people, depending on their own preferences of social and/or social consequences. For example, the cost of forming and integrating potentially contradicting beliefs into existing belief structures or risking conflict with one's in-group and social identity may outweigh the benefits of forming rich and accurate representations of others. Furthermore, the certainty associated with anticipating possible threats may outweigh the interest in learning more about unfamiliar people and social groups.

Finally, the study findings suggest that mental model change is uncomfortable and difficult to achieve, and therefore requires a strong motivation to invest the necessary cognitive effort. The results indicate it to be an uncomfortable process to challenge expectations, break down stereo-

types, and change representations of others. They also suggested, however, that this discomfort can be overruled by the desire to learn, establish more positive views of others, or establish a more accurate representations of others' complex human experiences. Weighing up and investing in different opportunity costs may therefore lead to very different outcomes of evaluations and mental models. The results above point to the potential benefit of including this insight in models of ToM.

Importantly, the findings at hand also suggest that modelling efficient and human-like ToM may not be possible without also modelling the errors or simplifications made by humans in social situations, such as considering information from stereotypes and inaccurate predictions. This poses the question of how accurate human ToM really is and what a model of ToM should strive for. Research has previously pointed to the limits of ToM (e.g. Keysar et al., 2003) but studies need to assess this question mechanistically, understand these limits, and establish whether it is indeed the goal to model ToM the way humans perform it naturally and intuitively. After all, the present results suggest big fundamental errors in the generation of human mental models. It may be possible to develop a better and still efficient alternative for AI.

Chapter 4

Theory of Mind Heuristics in Interactions

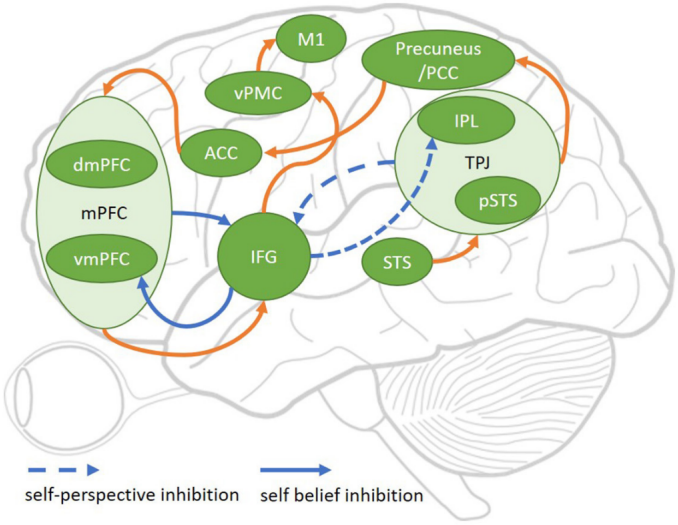
This chapter discusses two additional approaches to investigating what other cognitive strategies and mechanisms contribute to robust, efficient, and dynamic ToM. First, a study is discussed involving observation of an agent in a maze. Second, a study is presented that explores more complex interaction in a strategic game where the human subject's opponent is using different levels of ToM reasoning.

4.1 Cutting Corners in Theory of Mind

The study presented in this section was designed and carried out in collaboration with Jan Pöppel, University of Bielefeld, Germany. I would like to acknowledge and highlight his contribution to the design and development of the materials and experimental setup, as well as the analysis and interpretation of results. The study explores higher-level and lower-level cognitive ToM processes in a maze environment and aims to understand the role of egocentrism as a ToM heuristic. The workshop paper presented at the *AAAI Fall Symposium 2022 on Thinking Fast and Slow and Other Cognitive Theories in AI* was slightly adjusted for this thesis to improve readability based on the feedback kindly given by the workshop attendees.

4.1.1 Introduction

The space of social interactions and reasoning about others' minds is vast and there infinite possibilities of mental state inferences in any given moment. In line with the overall theme of this thesis, this study explores what biases and heuristics can narrow down these possibilities to reduce the energy required to reason about others. There are various features that have been found to affect ToM use. Studies of cultural differences in ToM suggest an element of individual learning and experience in shaping ToM use (e.g. Kobayashi et al., 2007). Furthermore, there is evidence that the activity of another person, as opposed to them being passive, facilitates perspective-taking (Freundlieb et al., 2015). Moreover, the nature of the other appears to be



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

relevant for ToM use, i.e. whether it is a human or an inanimate object (Samson et al., 2010). Research also suggests that mood and emotion are related to the extent to which an individual may be egocentric in their perspective (Todd et al., 2015).

Interestingly, psychological research suggests that humans do not even always infer others' mental states, even when it would be useful to do so. Keysar et al. (2003) suggest that egocentric reasoning is surprisingly common in social interactive tasks. In this study, the nature and manifestation of egocentrism is explored. Specifically, this work studies how the timing of an egocentric cue can affect ToM reasoning and induce biases that reduce the overall possibilities of mental state inferences.

This study explores the role of egocentrism in ToM by applying the dual-process informed framework by Zeng et al. (2020). They present a neuroscience-informed model, recognising the different brain areas that have been shown to be involved in ToM. Their research distinguishes 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 4.1). 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.

4.1.2 Rationale

Informed by the model by Zeng et al. (2020), the present study distinguished between previously established beliefs and momentarily available egocentric cues as a means of determining ToM inferences.

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

- Humans are complex and there are plenty of possible inferences to be made about others' mental states. What cognitive dynamics and heuristics affect the determination of specific inferences?

Zeng et al. (2020) 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. The results of this study explore whether the distinction between *belief* and *perspective* is helpful in modelling ToM. To dissociate higher-level *belief* and lower-level *perspective*, it was manipulated what information is visible at the time of potential ToM performance. The paradigm of this study involves an egocentric visual cue that indicates where a target object is located. The visibility of this cue was altered across conditions to explore the impact of immediate visual cues on ToM inferences.

Based on the different pathways proposed by Zeng et al. (2020), the following hypotheses are proposed for different types of cues:

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).

4.1.3 Methods

Participants

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.

Apparatus

The maze videos used in this study are created with the tool used by Pöppel et al. (2021). A virtual agent moved along a pre-determined trajectory through a 2D space, searching for a book (see figures 4.2-4.4). 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.

Design

The study had a 1x3 between-groups design with the independent variable *cue visibility* (no cue vs partial cue vs full cue). Figures 4.2-4.4 show snapshots of the different conditions.

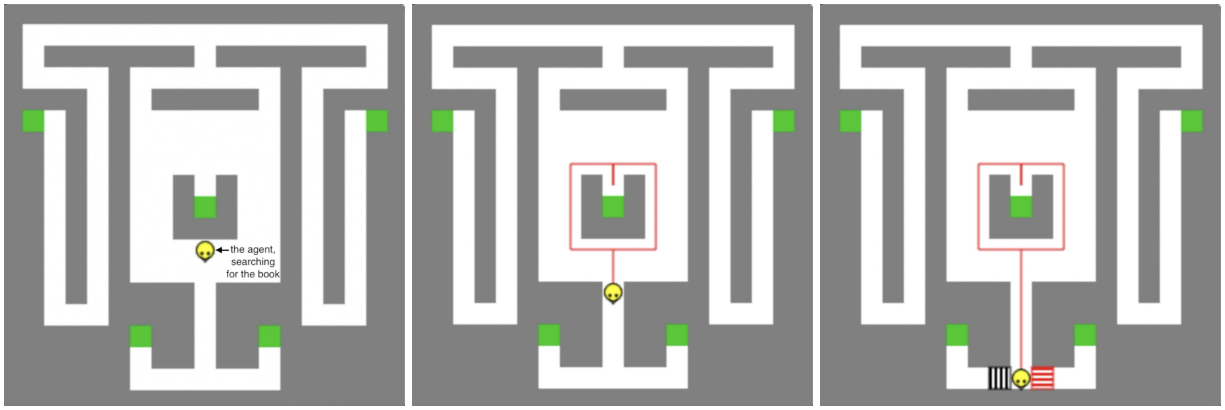


Figure 4.2: Condition 1, No Cue (participants do not see at any point where the book really is).

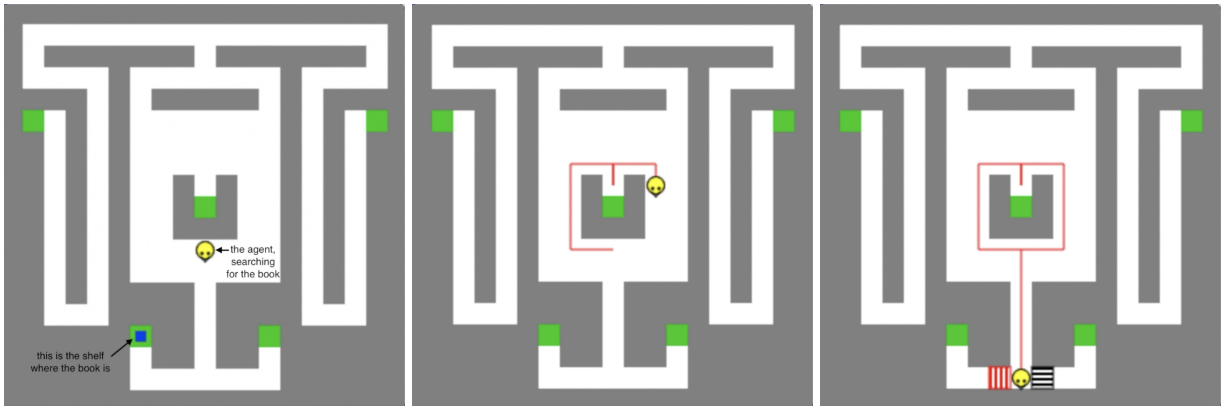


Figure 4.3: Condition 2, Partial Cue (participants initially see where the book really is but then it disappears, and the cue is not visible at the time of the prediction).

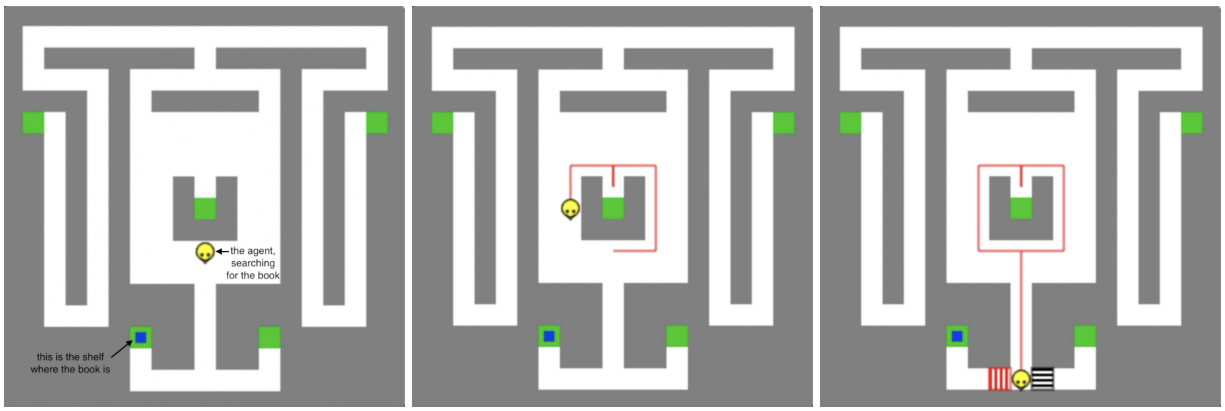


Figure 4.4: Condition 3, Full Cue (participants see where the book really is from the start to the end, including the time of the prediction).

The dependent variable *direction* was measured on a nominal scale with the categories *red* and *black*. Another dependent variable explanation was measured qualitatively by asking each participant why they chose their respective response. The qualitative element was included to illuminate not only the outcome of a person's ToM reasoning but also more about the type of information affecting the reasoning process.

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 4.5).

Figure 4.5: Answer choices.

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

4.1.4 Results

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

Quantitative Responses

Figure 4.6 and table 4.1 show which direction participants predicted the agent would 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.

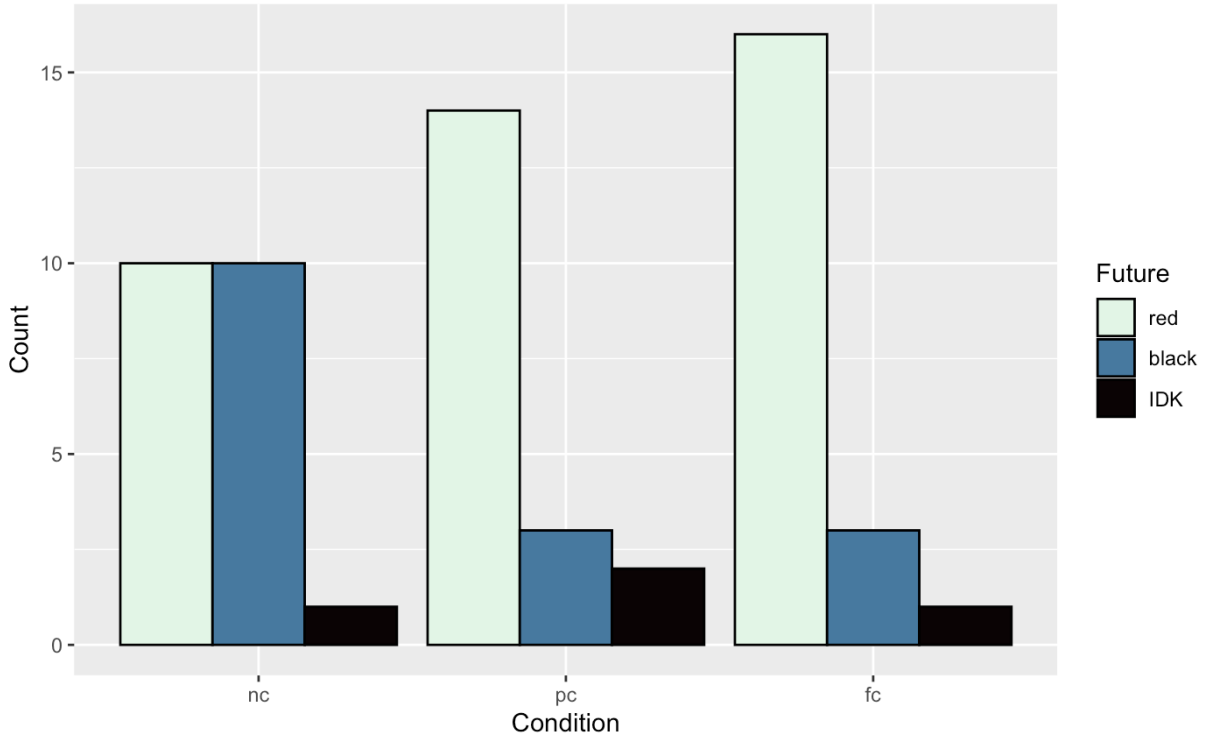


Figure 4.6: Counts of direction responses (Black, Red, & I don't know) by cue visibility (No cue, Partial cue, & Full cue).

A Chi-Square analysis was conducted and shows that the difference between the three groups is not significant, $X^2(4, N=60) = 7.76$, $p = 0.101$.

	Red	Black	I don't know
No cue	10	10	1
Partial cue	14	3	2
Full cue	16	3	1

Table 4.1: Counts of direction responses (Black, Red, & I don't know) by cue visibility (No cue, Full cue, & Partial cue)

Exploratory Analysis of Qualitative Responses

Participants were asked why they chose the answer they selected. As depicted in the framework by Zeng et al. (2020), the individual explanations were divided into categories based on *how*

relevant information was retrieved and *who* perspective was taken. Accordingly, each responses was divided into one of four categories (Table 4.2) based on the distinction between faster, more intuitive lower-level (perspective) pathways vs slower, more rational higher-level (belief) pathways, and self (egocentric) vs other (altercentric).

How	Perspective (fast) vs Belief (slow)
Who	Egocentric (self) vs Altercentric (other)

Table 4.2: Categories of ToM reasoning types.

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"

The analysis explored 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, $X^2(6, N=60) = 12.84$, $p = 0.046$ (Table 4.3).

	No cue	Partial cue	Full cue
Altercentric Belief	4	3	1
Altercentric Perspective	9	3	6
Egocentric Belief	0	3	5
Egocentric Perspective	2	7	8

Table 4.3: Counts of prediction strategies (Altercentric Belief, Altercentric Perspective, Egocentric Belief, & Egocentric Perspective) by cue visibility (No cue, Full cue, & Partial cue)

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, $X^2(2, N=60) = 0.45$, p

= 0.798 (Figure 4.7, Table 4.4). Interestingly, perspective-based reasoning was more common than belief-based reasoning in all three conditions.

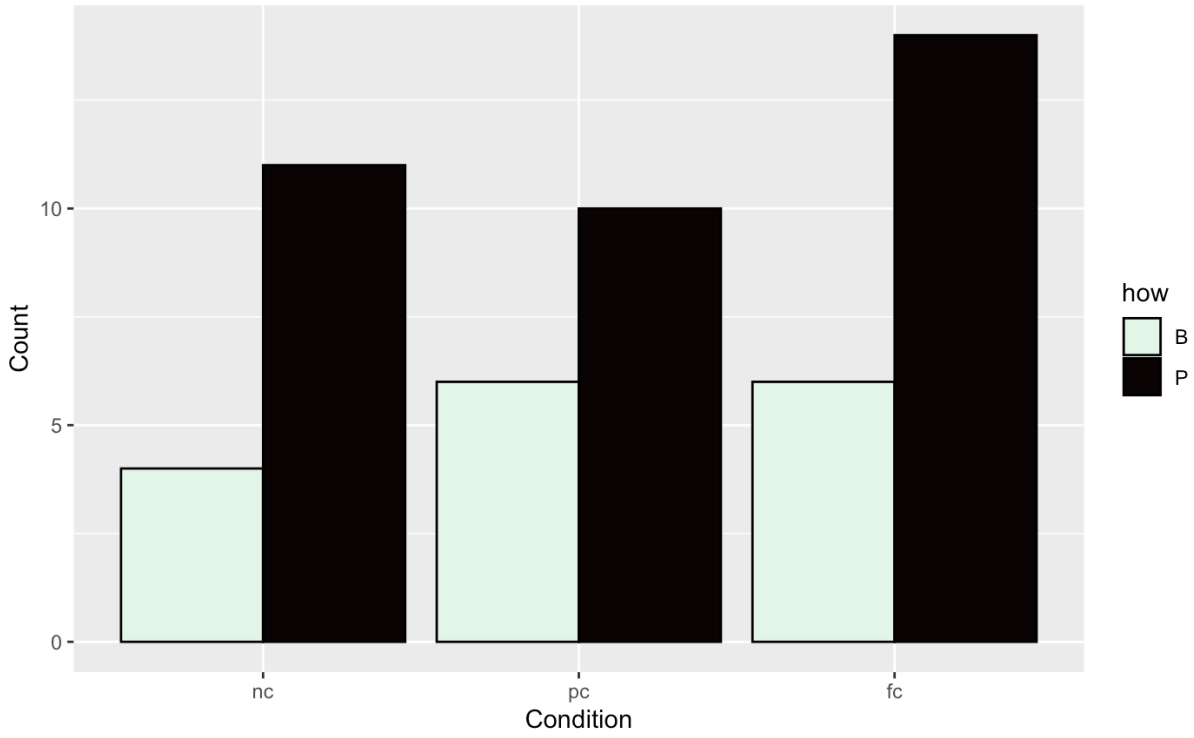


Figure 4.7: Counts of qualitative responses based on the source of information considered (belief vs perspective).

	No cue	Partial cue	Full cue
Belief	4	6	6
Perspective	11	10	14

Table 4.4: Counts of response type (Belief & Perspective) by cue visibility (No cue, Full cue, & Partial cue)

Participants did, however, differ significantly in what viewpoint they took depending on whether they saw a cue or not, $X^2(2, N=60) = 10.85$, $p = 0.004$ (Figure 4.8, Table 4.5).

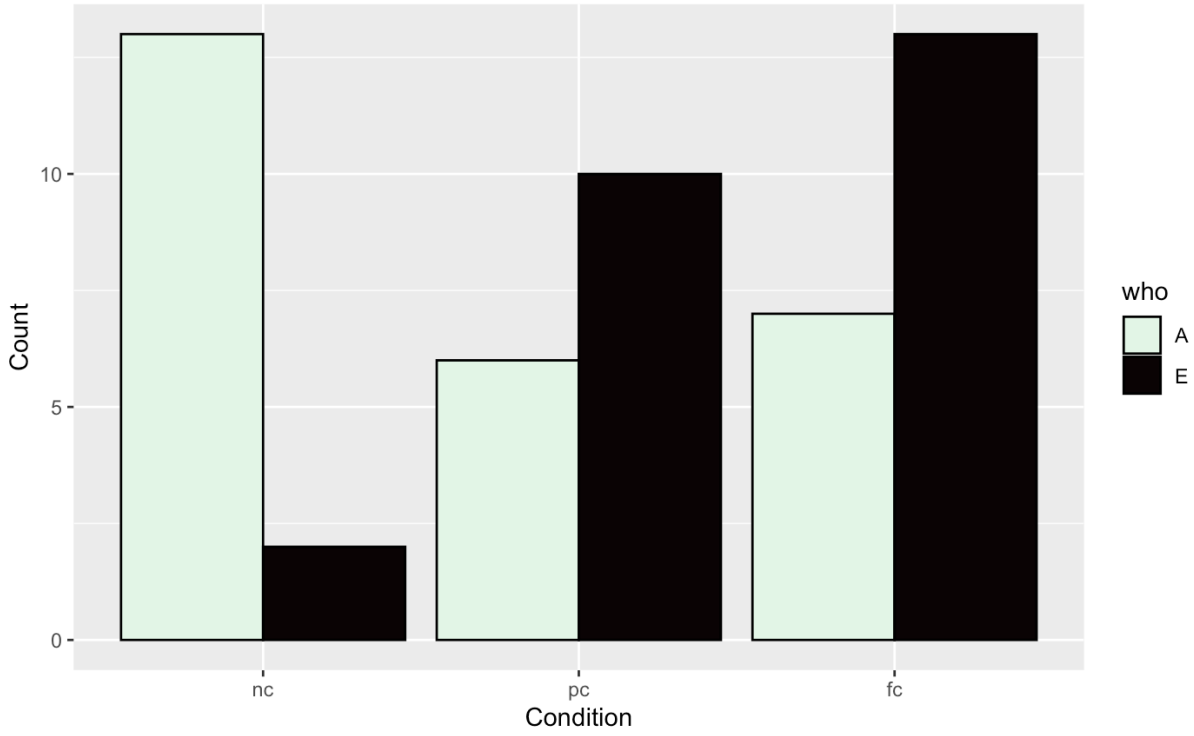


Figure 4.8: Counts of qualitative responses based on the viewpoint taken (own/egocentric vs other/altercentric).

	No cue	Partial cue	Full cue
Altercentric	13	6	7
Egocentric	2	10	13

Table 4.5: Counts of response type (Altercentric & Egocentric) by cue visibility (No cue, Full cue, & Partial cue)

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.

4.1.5 Discussion

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. Statistical analyses do not support the significance of these differences. However, the result is close to significance and possibly under-powered with only 20 participants in each condition. It is therefore suggested 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).

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. Perspective-based reasoning was more common than belief-based reasoning in all three conditions. Zeng et al. (2020) 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 previously argued that neither "fast" nor "slow" ToM can occur in isolation of the other (Jacob, 2019).

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. This suggests a high sensitivity to egocentric information in the context of social interactions, regardless of when it is shown, consistent with previous literature (Keysar et al., 2003). The results suggest a preference for "fast" ToM over "slow" ToM when an egocentric cue is available. It is noteworthy that the explanations of participants' choices differed significantly (by viewpoint considered, i.e. self vs other) 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.

Psychological Perspectives

ToM use can vary greatly 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 (Gelfand et al., 2011), and languages differ in the extent to which they include the vocabulary needed to communicate ToM concepts (Pyers and Senghas, 2009).

Rather than thinking about ToM as a binary ability, which is either present or absent, the present results show the possible variance in the factors contributing to mental model formation. The study at hand indicates the role of egocentrism as a cognitive tool that guides mental state inferences, both via current perspectives and previous beliefs. Visual cues had more obvious effects when still visible in the moment when a decision was made, but they also affected results when retrieved from memory. This highlights that both current and previous experiences can

strongly affect an individual's reasoning about another person by providing reference points and cues that shape new interpretations.

The results suggest that when an egocentric cues was available, it was very commonly drawn upon to inform a person's reasoning and decision. Many participants used their own tendencies or strategies to predict the agent's actions. It is proposed that egocentrism, i.e. self-related over other-related tendencies, can act as a heuristic to reduce the cognitive cost required to reason about others. Interestingly, there is also a tendency by participants to make at least some prediction rather than indicating that they didn't know where the agent would go. This study has made a case for the the value of qualitative rather than quantitative research to identify and illuminate these subtleties.

Modelling Perspectives

From a modelling perspective, the results of this study imply a conceptual benefit to including a distinction between self and other. A built-in bias towards the own perspective over other perspectives may be one way to formalise a heuristic contributing to efficient ToM, and to capture the behavioural observations in this study and previous work.

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 does not mean that it is flawless. Biases come at the cost of errors which are often very difficult to detect (Bartlett, 2017). 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.

4.2 Insights

Recall the research question for this chapter:

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

- Humans are complex and there are plenty of possible inferences to be made about others' mental states. What cognitive dynamics and heuristics affect the determination of specific inferences?

The results of this study offer the following explanatory hypotheses:

Theory of Mind Heuristics in Interactions

- ToM is commonly underused
- A way to reduce cognitive costs is using a model of the self to inform models of others

The distinction between perspective vs belief considered in this study fits into the overall concept of ToM by offering possibilities for heuristics in the element of perceptual elements, interoceptive signals, and mental state representations (highlighted in Figure 4.9). The results presented above illustrate instances where either of those elements can involve information about the self or the other person, where egocentric information is typically prioritised. As suggested in the discussion, it is likely not either previous beliefs or current experiences that affect decisions and behaviour, but a combination of both influences, creating dynamics of heuristics that are here hypothesised to increase efficiency in ToM.

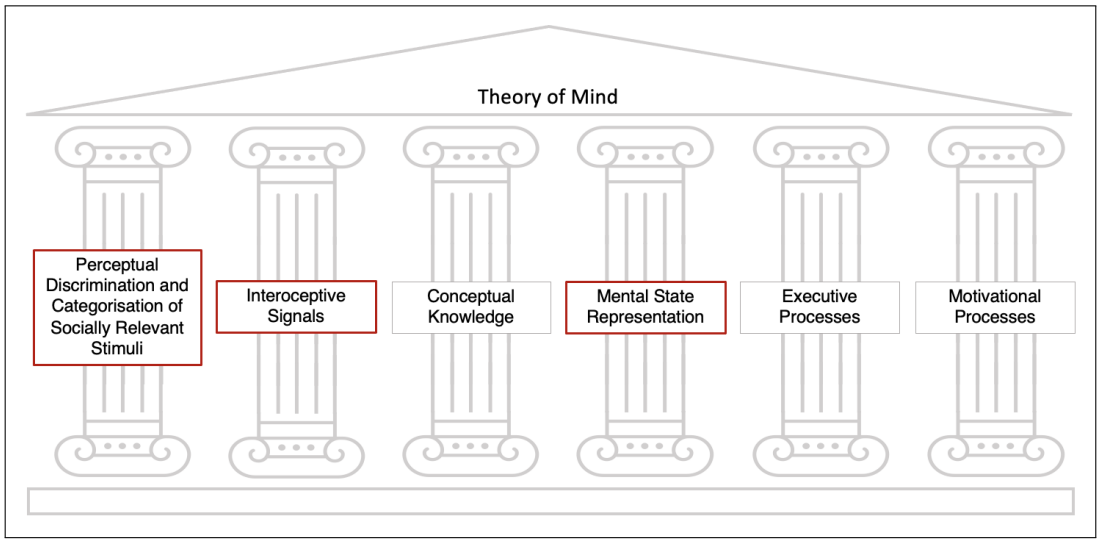


Figure 4.9: ToM elements considered in this thesis, based on Schaafsma et al. (2015).

Egocentrism is an interesting sub-component of ToM. In line with previous research, the findings at hand suggest an under-use of elaborate ToM and consideration of egocentric information to reduce processing costs. This involves a spill-over from what an individual knows

about themselves to what they would predict another person would think, feel, or do. That way, inference computations themselves could be kept simpler, and mental models could be generated faster. Egocentric ToM may, however, lack in accuracy and complexity compared to ToM that is primarily informed by observations of others' behaviour.

4.3 Theory of Mind in a Strategic Game

This study records and studies think-aloud responses in a strategic game scenario. It identifies different sources of information that can shape mental models and explore the role of situational demands and individual differences.

4.3.1 Introduction

Rusch et al. (2020) propose that ToM emerges and increases with higher degrees of uncertainty and interaction. A study by Goodie et al. (2012) on human ToM use in simple strategic games, in contrast, suggests more ToM during simple strategic games than the simplicity of the task may suggest. The extent of ToM use is therefore still to be researched further. To deepen the understanding and conceptualisation of ToM mechanisms, this study investigates whether humans really use as much ToM in a simple strategic game as Goodie et al. (2012) propose. This work explores the extent to which ToM is used and aims to shed light on the heuristics may be involved in ToM to facilitate selective and efficient mindreading.

Gallagher and Fiebich (2019) suggest that there are many factors that can affect the way in which ToM is performed and informed, such as context, previous observations, or social norms. However, research has not yet specified the concrete ways these strategies contribute to forming and shaping mental models. Specifically, this study is concerned with the reasoning processes about the task at hand and whether an participants' focus is on the other player's mental states, or more task-oriented. This work explores what is on a person's mind while they play a strategic game and what reasoning elements contribute to this process. To be as clear and specific as possible, distinctions between different levels of ToM will be considered. These are defined in the following and described with examples from the strategic game. The next section specifies a narrower understanding of the way in which ToM is applied in this research, followed by a specification of the questions and aims of this study. This is followed by the description of methods and results, and a discussion of the study findings. Finally, this work presents a model of the mechanisms at hand, followed by a general discussion and future directions.

4.3.2 Levels of ToM

Due to its recursive nature (Gmytrasiewicz et al., 1991), ToM can occur at different levels of depth. For example, one can reason about others' mental states, or about what another person

might think about one's own mental states, and so on. Using example quotes from the results of this study, the sections below describe the different levels of depth this study is concerned with, and how they are defined.

No ToM

Not using any ToM at all would be to simply reason about the world without considering others as agents: "I think I choose. [...] Next, yeah. Because I won't take just one point. That doesn't make sense."

Level 0 ToM

Level 0 ToM is defined as reasoning about others' actions but not considering their mental states, i.e. the possible reasons for their actions (e.g. goals, feelings, motivations, intentions, ...): "No, that doesn't make sense. If I move to B then logically, Papaya would end the round, so no. I'm going to leave with my three points."

Level 1 ToM

Level 1 ToM is understood as reasoning about another's mental states on a basic level: "I'm assuming they won't want that. So that would be, yeah [...] because at first I thought that they wouldn't want that because I would get more points but I would immediately go to D if I had the choice."

Level 2 ToM

Level 2 ToM is defined as a recurring instance of ToM and involves reasoning about what another person will reason about one's own mental states (or about another's mental states if the scenario was different): "[...] but I'm not willing to risk that because they know I'm not stupid enough to go for D."

Higher-level ToM

Beyond these levels, is possible to reason about mental states on even higher levels of recursion (theoretically infinitely) but this is less common in everyday scenarios (Gmytrasiewicz et al., 1991). In the present study even level 2 ToM was observed only very rarely. When it does happen, the computation of it should work in the same way as the change from level 1 to 2. For this reason, these higher levels of ToM are not explored here in more detail.

4.3.3 Rationale

Ultimately, any definition of ToM shapes the extent to and way in which ToM is recognised and identified as such from the collected and analysed data. Researchers have studied many different aspects of ToM under the same label (Schaafsma et al., 2015). Therefore, the importance of specifying what exactly on refers to when they use the term "Theory of Mind" is highlighted. Researchers have distinguished between explicit, slow, and deliberate vs implicit, fast, and automatic ToM (Apperly and Butterfill, 2009). Moreover, some literature is concerned with more emotional elements (Alfonso et al., 2015; Todd et al., 2015) and other work focusses on primarily cognitive ToM (Conway et al., 2019; Gweon et al., 2011; Wimmer and Perner, 1983). As in previous chapters, in this study the definition of ToM is restricted to explicit and cognitive manifestations. This includes the conscious understanding and processing of others' mental states as part of cognitive reasoning about another person in a situation of social interaction. Even within this understanding of ToM, there are various ways in which ToM can manifest across different individuals (also see chapter 4).

This manifestation of ToM requires many different cognitive resources, such as representing the other person's previous actions, inferring mental states, storing and accessing commonly encountered behavioural tendencies, and/or simulating possible future actions. As suggested by previous literature (Keysar et al., 2003) and explored in the previous study, ToM is commonly underused. This may be a way to reduce the overall spending of cognitive resources and energy. This study explores to what extent humans really use ToM in solving a simple strategic game and to what extent other reasoning strategies are employed. Beyond egocentrism, this study considers the prevalence of task-oriented over ToM-related reasoning. It has been suggested that there are many different cognitive strategies contributing to ToM, which can be used flexibly depending on the situation and the individual (Gallagher and Fiebich, 2019). This may again be a way to reduce overall ToM cost by employing a choice of less costly over more costly strategies. Insight into both these points of study are fundamental to better understanding ToM and creating a robust theoretical basis for a conceptual model. To investigate how much ToM (as defined above) is used and what the underlying mechanisms are, this study applies a strategic game based on the paradigm by Goodie et al. (2012) and investigates the following questions:

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

- Humans are complex and there are plenty of possible inferences to be made about others' mental states. What cognitive dynamics and heuristics affect the determination of specific inferences?
- What cognitive dynamics and heuristics characterise ToM variety?
- What cognitive dynamics and heuristics contribute to maximising ToM efficiency in social interactions?

4.3.4 Methods

Participants

This study was approved by the ethics board of the Psychology department at the University of Glasgow. G*Power was used for a Power Analysis and parameters were set to *Linear multiple regression: Fixed model, R2 deviation from zero*, with alpha set at 0.05, power at 0.8, the number of predictors at 2, and a medium effect size based on Goodie et al. (2012). This G*Power calculated a sample size of 23. The Psychology department's participant pool was used to recruit 40 participants who were all students at the university.

Design

This study has both quantitative and qualitative elements. For quantitative measures, each participant played the strategic game against both a myopic agent (not reasoning ahead and therefore not playing optimally) and a predictive agent (reasoning ahead and playing optimally). Participants were evenly divided to two conditions: *MP* (myopic agent first, predictive agent second) and *PM* (predictive agent first, myopic agent second). Thereby, each game followed the same sequence of point distributions. All video and audio was recorded and later re-coded to measure the dependent variables *points* (the total score indicating game performance), *future* (the extent of reasoning about future turns), and *ToM level* (the level of depth in reasoning about others' mental states). These variables were measured on a ratio, categorical, and categorical scale, respectively. The qualitative element of this study is a think-aloud component (Charters, 2003). Each participant was asked to verbalise their thought processes in a stream of consciousness format, while playing the game. Participants were encouraged to say as many thoughts aloud as possible. Think-aloud paradigms have been proposed to be a valuable method to study continuous thought processes (Charters, 2003).

Materials

The strategic game utilised in this study (example in Figure 4.10) is based on the paradigm by Goodie et al. (2012).

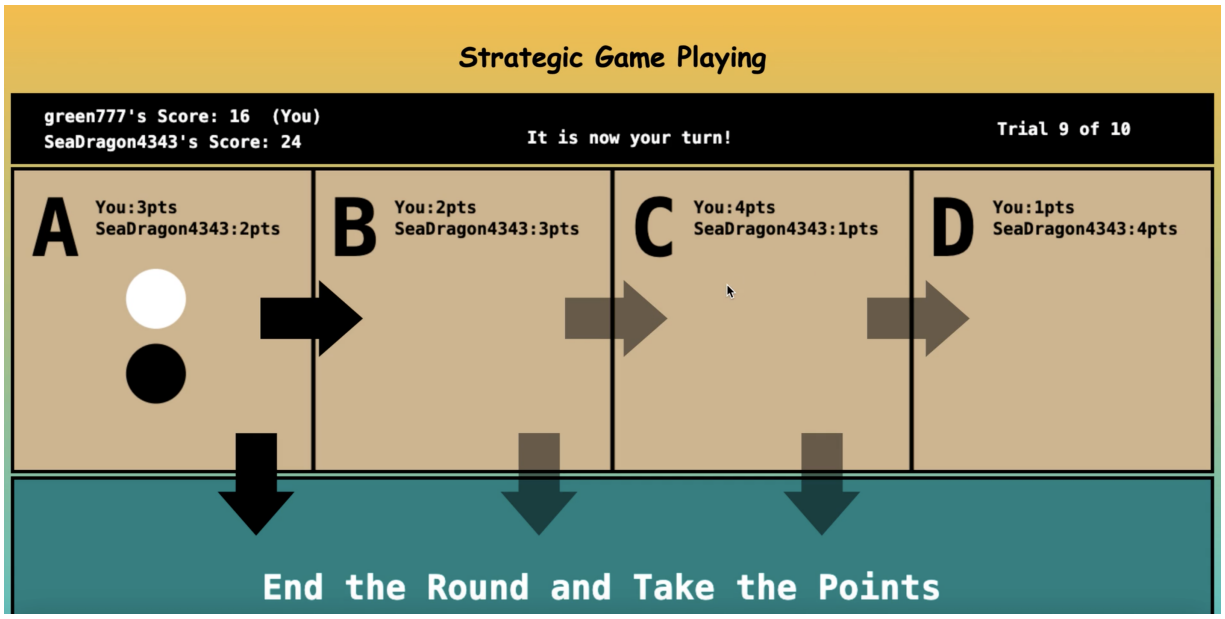


Figure 4.10: Example from the strategic game.

The game is interactive and turn-based, with the goal to collect more points than the opponent. Each game consists of 10 rounds against the same opponent and there is no time limit to taking turns. The players take turns in starting the rounds and both players always move together. When it is a player's turn, they can decide to either move both themselves and their opponent to the subsequent box (click on right-arrow) and it is then the other player's turn, or to end the round (click on down-arrow) and both players earn the points from the box they are currently in. The points in the four boxes are 1-4, 2-3, 3-2, and 4-1 (always adding to 5). The distribution of these points in the different boxes varies from round to round but is always displayed for the full round. The overall sequence is the same for each full game (i.e. 10 rounds) and fair for both players. The myopic agent is programmed to only consider the current and subsequent box when making a decision. If the current points are higher than the points in the next box, it ends the round. If the points in the next box are higher than the points in the current box, it continues. The predictive agent, in contrast, reasons one more step ahead. If the next box has higher points than the current box, but the second next box in the sequence is even more beneficial for the opponent, it will also exit in the current box. In the present study, each participant played one game against the myopic agent and one against the predictive agent, in a randomised order. The game was played on a laptop to allow recording of audio and all responses on screen.

Procedure

Participants were recruited online and invited into the lab to play the strategic game. After arriving, they gave informed consent. The researcher then explained how the game worked and instructed participants to speak aloud while playing it. The agents were referred to as "opponents" and spoken about as if they were another human in a different room. The researcher then started the game and left the room. Participants completed the tasks and let the researcher know when they were finished. Before leaving, they were fully debriefed and received a payment of £10 for their participation.

4.3.5 Results

Results were initially screened for completeness. 6 participants were excluded from the analysis due to missing data and/or poor verbal responses. The analysis below is therefore based on a total of 34 responses, which is still above the G*Power estimate of 23. The available data was analysed with regards to the extent of ToM use, primarily quantitatively, and with regards to efficient ToM strategies, primarily qualitatively.

Quantitative results

The dependent variables *points* (the total score indicating game performance), *future* (the extent of reasoning about future turns), and *ToM level* (the level of depth in reasoning about others' mental states, as defined above) were manually quantified from the recordings. Thereby, *ToM level* ranges from 0 to 2 and *future* ranges from 0 (the box they are currently positioned in) to 3 (box D, if they are currently in A). For both *future* and *ToM level*, NA values marked that there was no indication of ToM depth and extent of future reasoning.

Descriptive statistics

Figure 4.11 shows the distribution of total points scored in both conditions, against both opponents.

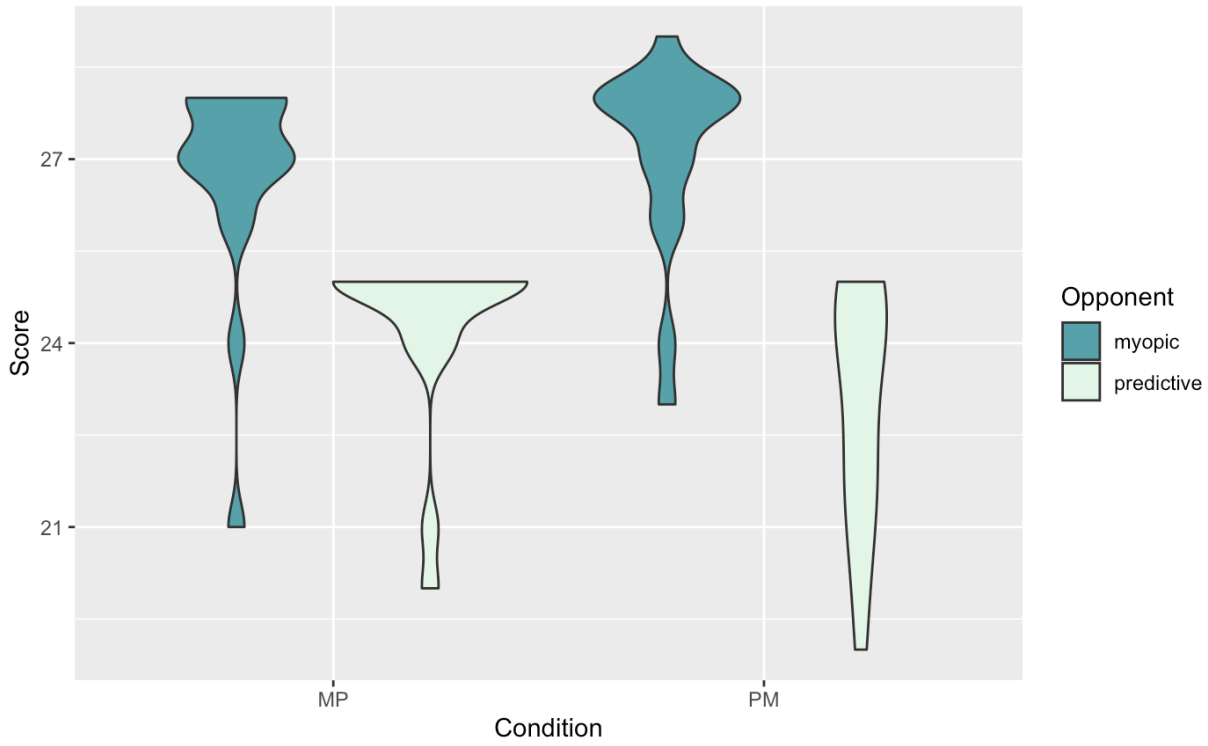


Figure 4.11: Total points by opponent in both conditions. Recall, the myopic agent does not reason ahead. The predictive agent predicts optimal choices and makes optimal decision itself.

Figure 4.11 shows overall higher scores against the myopic opponent than the predictive opponent. The figure also indicates an interaction effect, with scores across the two opponents diverging more when participants played against the predictive agent first and the myopic agent second. Figure 5.12 captures all counts of recorded ToM depth and reasoning into the future, as annotated from the think-aloud recordings.

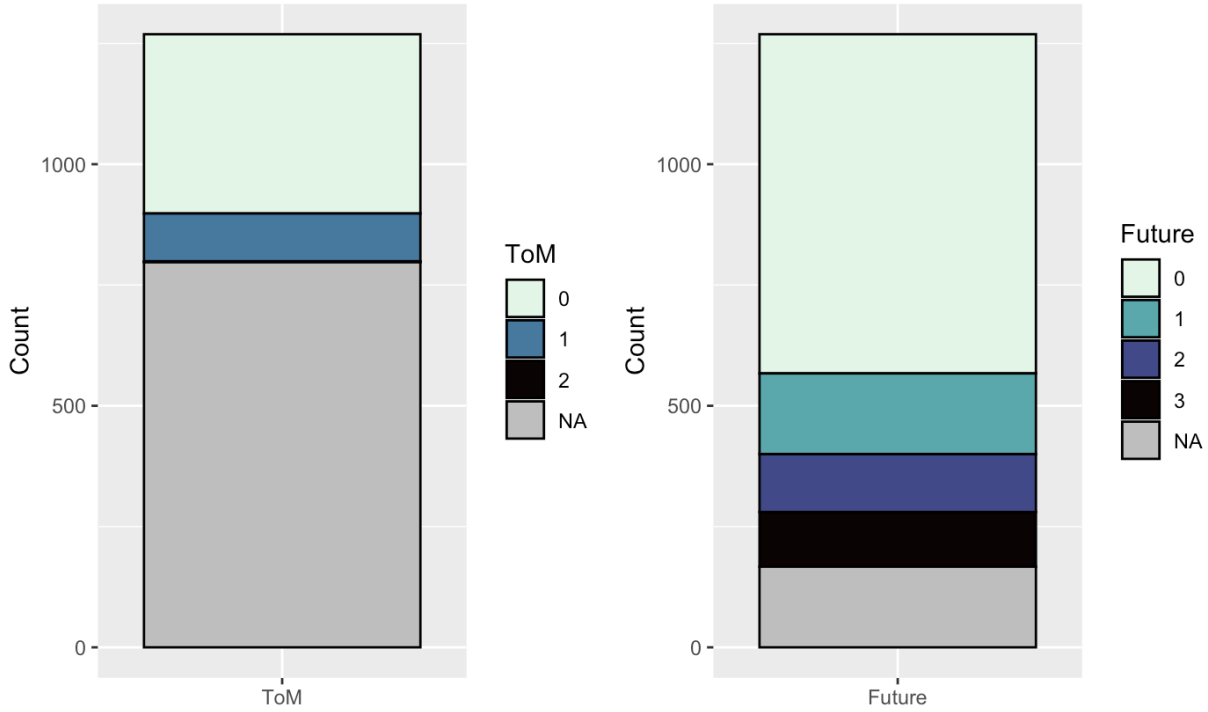


Figure 4.12: Total counts of ToM depth and reasoning into the future, as quantified from the recordings.

The graphs indicate much lower values of evident ToM at any level compared to level 0 or no available indicators of ToM. Similarly, reasoning into the future made up only a small proportion compared to reasoning only about the current position or not giving an indication of how far ahead is being reasoned. Figure 4.13 represents the association of recorded ToM depth and scored points.

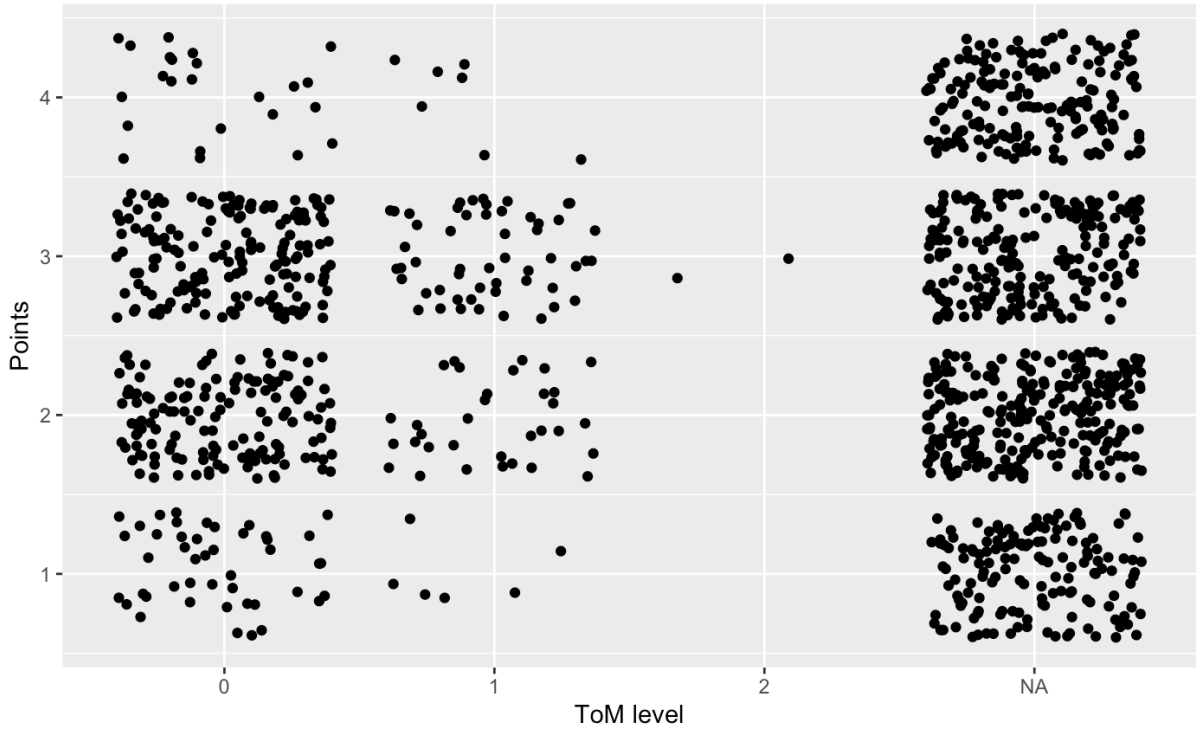


Figure 4.13: Recorded scores by recorded ToM depth.

The figure suggests no differences in scores by ToM level, which indicates that there is no association between the two variables. Rather, the figure reflects differences in total recorded data points across ToM levels. Most verbal responses did not give any indication of ToM reasoning (NA). Of the ToM-related responses, most were at level 0 of ToM. Figure 4.14 shows the association between scored points and recorded level of reasoning into the future.

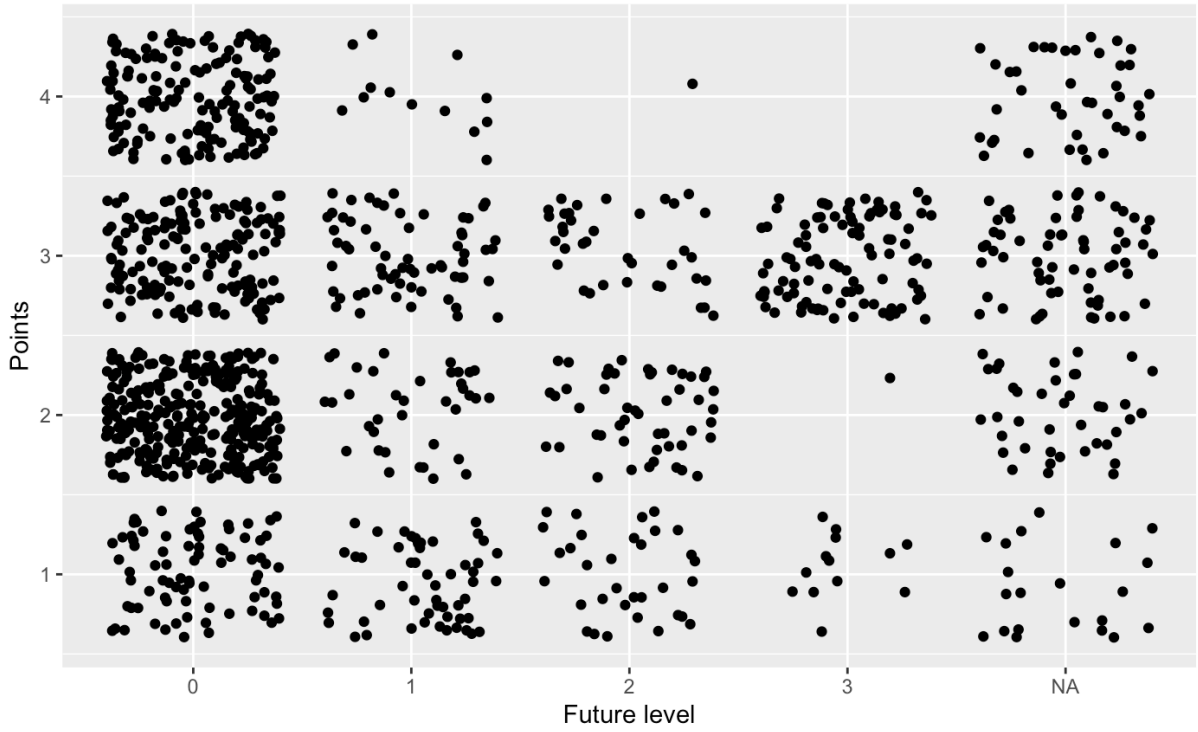


Figure 4.14: Recorded scores by recorded future level.

Figure 4.14 does show quite different patterns of scored points by recorded future reasoning, suggesting that there may be a relationship between the two variables, with higher scores as future reasoning increases. The following section will include analyses to investigate whether the relationships between 1) *condition* (i.e. order of opponent), *opponent*, and *score*, 2) *ToM level* and *points*, and 3) *future level* and *points* are significant.

Inferential statistics

A multiple linear regression analysis was conducted to assess the effects of condition and opponent on participant scores. The analysis (adjusted R^2=0.49) showed a non-significant relationship (p=0.485) between condition and score (B=0.41) and a significant relationship (p<0.001) between opponent and score (B=-2.41), reflecting a 2.41-point decrease against the predictive opponent compared to the myopic opponent. The analysis also revealed a significant interaction effect (p=0.044) between condition and opponent on scores (B=-1.71).

The same analysis was also conducted excluding the first 4 trials to assess whether it would make a difference to remove initial rounds and give participants the time to establish mental models of their opponents. This analysis (adjusted R^2=0.24) revealed a non-significant relationship (p=0.475) between condition and score (B=0.29), and a significant relationship (p=0.048) between opponent and score (B=-0.82), and an only just non-significant interaction effect (p=0.089) between condition and opponent on scores (B=-1.00).

To account for the categorical nature of the variables ToM level and future level, the two

variables were each re-coded with deviation coding. Firstly, a simple linear regression analysis was carried out to assess the relationship between points and ToM level, with the deviation coding for ToM level based on level 0 as baseline. This resulted in non-significant results for all levels (adjusted R^2<0.001): 0 vs 1 (B=0.17; p=0.110), 0 vs 2 (B=0.56; p=0.409), and 0 vs NA (B=0.08; p=0.174). Secondly, a simple linear regression analysis was conducted to investigate the relationship between points and level of future reasoning, with the future reasoning level deviation coding also based on NA as baseline. This analysis (adjusted R^2=0.05) revealed significantly lower points compared to level 0 for level 1 and 2, and significantly higher points compared to level 0 for level 3 and NA: 0 vs 1 (B=-0.34; p<0.001), 0 vs 2 (B=-0.51; p<0.001) 0 vs 3 (B=0.20; p=0.031), and 0 vs NA (B=0.20; p=0.014).

To summarise, the present data suggests that participants scored better against the myopic than the predictive opponent. The order of opponents did not affect overall results. However, there was an interaction between the two variables: The difference between scores against the myopic and predictive opponent were closer together when participants played against the myopic agent first, and diverged more when they played against the predictive agent first. The same trends were observed when the first 4 rounds were excluded from the analysis as a trial period to allow participants to observe what their opponent's strategy might be. There was no difference between recorded ToM levels in their relationship with scored points. Recorded future reasoning at level 0 was associated with significantly lower points than level 3 and NA, but not levels 1 and 2. In the following section, qualitative results based on the collected think-aloud data are presented and all results are subsequently discussed.

Qualitative results

As in studies 1 and 2 of this thesis, the AMEE Guide No. 131 (Kiger and Varpio, 2020) was followed to analyse qualitative responsive of the present study thematically. The six steps of 1) data familiarisation, 2) initial code generation, 3) theme search, 4) theme review, 5) defining and naming of themes, and 6) reporting the results, were followed as recommended by the authors of the guide. The resulting findings are be presented in the following, laying out three themes and their respective sub-themes.

Theme 1: *Individual differences in ToM use*

The results of this study suggest large individual differences in ToM use, with regards to both quantity and quality. The following sub-themes illustrate various grounds of individual differences observed in the data at hand.

i. Quantity and depth
   Findings included large differences in the extent to which participants reasoned about their opponent's actions. For example, some people considered their opponent's reasoning and

actions deeply and throughout playing the game. "So if I move over then they are going to take it out there, [...] which means they'll be a point ahead of me [...] but I'm not willing to risk that because they know I'm not stupid enough to go for D." "Yeah, you should take too. Yeah, yeah. [...] That's it. He played a lot better, I think." Other participants considered their opponent's actions more superficially "So he gets 3, I get 2 in this one. [...] In the next one I get 3, he gets 2. [...] Wait if I'm, if I move on, he's going to move on as well and get 4 points. [...] So, going to take the points now. So we're even, 10 - 10." or solely focussed on their own points. "OK, if I choose this [...] I will only have a one point deficit [...] but if I go to the next one and then I will have a 3 point deficit" The analysis suggests that the amount of recorded ToM appeared neither necessary nor sufficient for participants' success in playing the game. The data shows evidence of all patterns, scoring high with high ToM use, scoring low with low ToM use, scoring low with high ToM use, and scoring high with low ToM use.

ii. Demand-driven & ad hoc ToM

Participants' use of ToM generally increased when they were required to make a decision because it was their turn, and decreased when it was the opponent's turn. For example, one participant reasoned less deeply about their opponent's situation and next actions "OK, this could be good because it will probably move along where it'll loose and then I get to move on. [...] Or it might take it cause it will know that I'll get 4 points" and more thoroughly when faced with the exact same decision right after. "OK, so let's move it along. [...] See the problem with this one is, if I go out in this round, I know that I'll lose. [...] But. And then the scores will be equal. [...] But if I move along, it will move along again, and then I'll lose by more so that would be stupid, so I'm going to. [...] End the round here." In a similar fashion, it was common for participants to predict the opponent's actions and think through the whole problem again when faced with the same sequence shortly after. "something tells me to end here but let me check"

iii. General vs specific mental models

Participants differed greatly in the extent to which they registered the behavioural differences across agents. Many did not comment on any differences at all and reasoned about both players more generally as their opponent "It's better to [...] lead by even just one rather than take a lot of risk [...] because you don't know what the person is going to do" whereas others clearly noticed the differences between the two players. "OK, I'm glad I've got someone who's... she's no Sea Dragon." "Oooohhh. [...] So papaya's mindset is really different from the first player." The results also show a considerable extent of individual differences in the sources of information that were used to infer others' mental states or predict their actions. This will be presented in detail as theme 2.

Theme 2: ***Sources of information***

The explanations given to support participants' reasoning and predictions about their opponent's actions differed greatly across individuals and situations. Five categories of information sources were established and are presented below.

i. Other's behaviour

A common type of explanation for predicted behaviour or judgement was the use of observed past behaviour. "So they decide to move and I'll naturally decide to move. [...] So they're a bit more risky. [...] They don't mind taking the chance." When observed, these insights would often be applied to participants' choices and strategies. "Uhm, so I've got 3, they've got 2, but I can end up with one. [...] 3 if I will, they'll take it. [...] Or they might be tempted and take, [...] and try moving in the four points [...] Like they did before, if I think about that. [...] And they've shown to be quite risky." Predictions could also cross over from one agent to the other, for example by playing more conservatively against the second opponent in response to mistakes against the fist "OK, learning from my mistakes. I will exit here just so that I am sure I will get more points than my opponent." or by recognising a good strategy and continuing with it against the next opponent. "So he usually ends in the first, [...] in the first box. Oh my God. [...] I know how to win. [...] Yes [...] Just end [...] Yes [...] Yes, I know how to win it."

ii. Egocentrism

Another influence on participants' reasoning and mental models was egocentric information. Thereby, individuals applied what they knew about themselves to their opponent, for example by predicting other player to play with the same strategy as themselves. "The other player will. [...] Take those points because they'll see that I would take us to the end." One participant was surprised by the opponent's actions, "My opponent will most likely [...] choose to move to the third spot after. [...] OK. [...] Huh, that was unexpected." but then proceeded to doing the exact same moves in the next round under those same circumstances. Another concluded that executing the same actions meant that there was the same underlying reasoning. "OK, so we're both tied and we both had the same thought process, OK"

iii. General mental model

A third way of informing mental models that was observed was to use more general previous knowledge and expectations and apply them to this situation. For example, many participants had assumptions about their opponent before the game started and any behaviour could be observed. "I might have [...] overestimated you papaya [...] Well, that's great. [...] So maybe I was under the wrong assumption that [...] everyone was very intelligent." Interestingly, the vast majority referred to their opponent as "he". "but maybe he's going to decide to move on, [...] but he's never going to do that because [...]

the C square is 4 for me, 1 for him, [...] and he knows that I'm going to, you know, take the points [...] and not move on to the last one." The default for many participants was also to expect their opponents to play optimally. "So you know, if I thought they were stupid, maybe I would risk more. [...] But I think they might make the most optimised choice. [...] So you know, I'm sticking with what they do." It was even common to notice poor choices but still maintain the expectation of subsequent optimal responses. "No, that doesn't make sense. [...] If I move to B then [...] logically, Papaya would end the round, so no. [...] I'm going to leave with my three points."

iv. Game logic

"It is getting easier. [...] When you understand. [...] The law of the game." Another factor influencing participants' choices and reasoning was a focus on the logic of the game itself, independent of behavioural tendencies or previous actions. Many planned several moves into the future and reasoned about their opponent's decisions according to what would be the ideal response in that situation. "It's 4 to 1, so if I move on to the next, I think he's going to take the points. [...] Like, the good strategy for him would be to take the points." "If I go on to C, [...] I get 3 points, they get 2 points, but chances are they'll go on to D [...] which means they'll have a round of 4 points, so it's logical for me to end at B."

v. Goals

A final way of informing thoughts and decisions was to directly act on the own goals. This type of reasoning was less deep and more direct, not considering game dynamics or different alternatives. "I'm still losing, but I think I want to get 4 points and they get 1 point, so."

Theme 3: ***Using mental models***

This theme is concerned with the practical application of mental models. Generating a mental model from different pieces of available information did not manifest in behaviour in a straightforward fashion. Rather, more complex patterns of the use of mental models of participants' opponents were observed in practice.

i. Observations vs behaviour

Interestingly, observing opponents' actions and predicting their subsequent choices as either good or bad did not necessarily translate into behaviour. For example, even when participants observed their opponent's behavioural tendencies as non-optimal, many decided to avoid taking risks. "And they've shown to be quite risky. [...] So, ok, let's see.[...] Still two point difference, so I should play it safe if I can. [...] So that's me playing safe again." "Oooohhh, she, Papaya choose to move to block C. [...] [next game] So I choose to quit [...] at stage A for safety."

ii. Risk-taking

In general, maintaining the same strategy independent of agent was more common and referred to as "safer" than changing the own strategy in response to the opponent. "It's better to [...] lead by even just one rather than take a lot of risk [...] because you don't know what the person is going to do." Many participants defaulted to taking three points whenever possible, independent of their opponent. "So I'm, it's my turn, so I'm going to take the three points [...] because I need to gain some more points." "Yeah, because he's not going to give me. So if I move on, [...] he's not gonna let me get all the way to the two and three points. [...] So it's probably better if I just take them early."

iii. Confidence

Even though many participants acted on the mental models of their opponents in a primarily conservative way by "playing safe", the predictions of future choices were expressed with consistent confidence. "They'll definitely collect it from there. They'll collect two points. Give me three points" "They won't move [...] Yeah" For example, one participant very confidently believed that the second opponent would make good decisions before the game started even though their first opponent played poorly. "I'll have to move to block B [...] and take that chance. They'll definitely collect it from there."

4.3.6 Discussion

Both quantitative and qualitative results are discussed below with regards to previous literature.

Quantitative results

The first main component of the present quantitative analysis was the association of overall scores with the opponent (predictive vs myopic), the order of playing against the two opponents, and the interaction of the two.

The quantitative analysis of this study revealed that the opponent but not the order of playing the two opponents significantly affected participants' overall score. The finding that participants scored better against the myopic opponent than the predictive opponent is an intuitive result because the myopic agent does not consider future scenarios and it is therefore generally easier to score highly against it than against the predictive agent. The order of playing the two agents did not predict overall scores. This may initially suggest that there is no deployment of a unique mental model for each opponent, as the process of learning about the game in the initial rounds did not seem to be affected by which opponent participants played against first.

The significant interaction between the two variables shows that the order of playing against the two opponents affected the extent to which the scores differed across opponents. Scores diverged more across the different opponents when participants played against the predictive agent

first and less so for those participants who played against the myopic agent first. Research suggests that learning involves the formation of biases which shape responses in future situations (Greggor et al., 2017). Participants who played against the predictive agent first played a lot better against the myopic agent in their second game. Participants who played against the myopic agent first did not have the advantage of learning from a previous game and played worse against it than participants who played against it in their second game. They, in turn, played better against the predictive agent than those participants who played against the predictive agent first. It is likely that all participants improved their strategies as the game went on. This learning pattern reflected in the interaction may indicate a difference in how mental models affected this improvement.

However, it seems that in both conditions participants' second performance was better compared to those participants who played against that agent as their first game. This suggests that it may not necessarily be the specific mental model that affected performance, but more a familiarity and experience with the game overall. It is therefore difficult to be conclusive about the use of ToM and its effect on participant scores.

The findings above were slightly weaker but similar in their pattern with the first 4 trials removed from the analysis. This is to be expected considering the availability of fewer data points. More importantly, however, participants did not seem to adjust to the new agent even when the first few rounds are not considered in the overall scoring. This is interpreted as an indication that biases are formed initially and then maintained throughout the rest of the game. The results do not, however, suggest whether this bias would be "located" in a mental model or in a more general game strategy: It is not evident from the score of points whether a certain mental model was developed and not updated despite new evidence, or a certain sequential strategy was developed and not updated despite new evidence, or both.

On the other hand, if ToM is being used consistently during this task, the phenomenon of initial bias formation and lack of updating could be one mechanism to contribute to more efficient processing within ToM specifically, instead of generally. Rather than forming a new model from scratch for a new person, it may be the case that as much information as possible is applied from previous observations with other people in similar situations. Especially in a social situation characterised by both uncertainty and interaction (Rusch et al., 2020), this application of independent but helpful mental models would be a possible way to fill gaps, reduce uncertainty, and guide complex interactions with others.

A disadvantage of the present analysis is the amount of recorded NA values (Figure 4.13). There are more data points of no ToM than any ToM but even more often there was no verbal indication of whether participants did or did not employ ToM. Future study of precise measures that can more certainly confirm or rule out the use of ToM in social interactions is therefore encouraged.

Interestingly, Goodie et al. (2012) found that participants who played against the predictive

opponent scored much higher overall than those who played against the myopic opponent. The present study does not show this particular difference in overall scores between the two agents, instead it suggests the opposite difference. However, participants in Goodie et al. (2012) had a much longer trial period and it is possible that the individuals who played against the myopic agent eventually disengaged.

The second main component of the present quantitative analysis was the association of individually scored points with recorded ToM levels and recorded future levels. Overall ToM recordings were low, and the majority of recorded levels of future reasoning was either level 0 or level 1.

Regression analyses resulted in non-significant relationships between collected points and recorded ToM levels. In their study, Goodie et al. (2012) concluded that higher overall scores were due to more ToM use. The present data, in contrast, does not support such a link between ToM use and scores. The ToM-based learning about the opponent's strategy that Goodie et al. (2012) refer to as only inferred from their points. However, there is no insight into the underlying reasoning. The present data indicates that other, non-ToM processes, such as strategic reasoning or habits, may be used more than Goodie et al. (2012) postulate. The finding that participants could score high or low independent of the ToM recorded suggests that ToM was applicable to this task but not required to achieve good results. In agreement with this possibility, Burge (2018), for example, claims that not all actions that may suggest ToM do really require it, and therefore do not necessarily indicate ToM use. For future studies it is recommended to measure ToM itself rather than possibly but not necessarily related outcomes such as performance scores in strategic games.

Furthermore, it was found that collected points were significantly lower for future level 3 and NAs compared to level 0, but not for levels 1 and 2. In other words, both level 3 reasoning and NA values were associated with higher points. Reasoning 3 boxes ahead is only possible in the beginning of the game. Firstly, this suggests a strategic benefit of thinking ahead and considering later-on point distributions, independent of ToM. Secondly, even when participants did not verbalise explicitly of how far into the future they were planning (NA), they sometimes scored higher compared to when participants were thinking ahead. These individuals might have still done the same reasoning without explicitly expressing it. Here, it was the case that NA values for future reasoning could still be associated with higher scores but NA values for ToM were not.

A limitation of the present way of quantifying the extent to which participants reasoned into the future and into the depth of their opponent's mental states was that only the words participants said aloud were accessible. They were asked to verbalise their reasoning as much as possible, and many participants spoke very consistently and thoroughly. However, there is always a chance that certain elements of their reasoning were not captured in the data. It is also necessary to highlight again that in this study the focus is on explicit ToM. It may well be that

there are other, particularly implicit, effects of ToM on decisions, that are not captured here.

Qualitative results

Before the qualitative results are discussed, it is essential to note that some participants did not think their opponent was a real person. However, this did not necessarily impact the extent to which they reasoned about possible mental states, strategies, and planned future actions. Mental models of humans may offer reference points to represent and reason about AI (Guzman and Lewis, 2020; Reeves and Nass, 1996). Many participants who pointed out that they thought they were playing an AI, did indeed still refer to their opponent's intentions, goals, or strategies, and call it "he" or "them".

The second research question asks what cognitive mechanisms contribute to ToM variety. Gallagher and Fiebich (2019) have suggested that humans have many different strategies available to perform ToM. The first and second theme show more specific elements contributing to ToM variety.

Firstly, theme 1 of the qualitative analysis (*Individual differences in ToM use*) captures a range of individual differences in ToM use with regards to the quantity and depth but also the situations that facilitated ToM and the specificity of the mental models themselves. For example, some people used no ToM at all, some showed ToM at higher levels and others with more recursive depth. It appears that the extent to which the different levels were used, decreased with the level of depth. The observed differences in recursive depth of ToM are consistent with the previous suggestion that at a certain level of depth, ToM performance does not increase anymore (Gmytrasiewicz and Durfee, 1995). Even beyond the question of benefit associated with different levels of depth, many instances of playing the game did not involve any ToM at all. Interestingly, when participants were in a more active situation because it was their turn to move and make a decision, the reasoning about their opponent was more frequent and more detailed. This suggests an influence of situational demands on the extent of ToM use.

Secondly, participant data showed large differences in the level of detail that characterised the reasoning about the other player. Some participants were very attuned to the opponent's specific behaviour, others treated the two players almost the same, more generally based on their role as the opponent. Beyond situational demands, there may be individual differences in the detail with which mental models are generated. It has been argued for less detailed models to control computational costs (Pynadath and Marsella, 2007), however the present data suggests that there may be differences in levels of detail beyond the bare minimum. Specifying the extent of this variety requires further research into the level of detail in mental models.

Thirdly, theme 2 (*Sources of information*) outlines a range of possible of sources from which information about others' mental states can be retrieved. Observations of others' behaviour, egocentrism, general mental models, game logic, and goals were identified as possible entities or structures that can inform the generation of mental models. A mental state representation

based on egocentric information (based on the person's own experience) may be very different to one based on observations of another individual's behaviour or expectations about what people would generally do. These different paths to mental models, based on what information is available and what type of source is commonly considered, can create very different inferences about others, and ultimately facilitate great variety in ToM.

Finally, the third research question of this work examines what other cognitive strategies participants used to maximise ToM efficiency. Theme 3 (*Using mental models*) indicates that observation of others' possible mental states does not necessarily mean that those observations affect an individual's actions. Even if another person's mental states are reasoned about and represented, this does not mean that this mental model always translates to behaviour. This highlights the notion that the strong conclusions made by Goodie et al. (2012) regarding ToM use underlying certain actions may be premature and the link between ToM cognition and action may not be as strong as they suggest. Rather, the data of this study indicates that observations of the other player's tendencies may be made in one moment but only at another point in time affect behaviour and changes in responses, when multiple observations add up to affect existing patterns and biases.

4.3.7 Psychological Perspectives

To summarise, the study suggests tendencies of minimal and selective ToM use: The reasoning in this competitive task may have been possible without consistent and deep ToM. Furthermore, the data at hand illustrates various possible individual differences in ToM. On a theoretical level, this illuminates the dynamic combination of different elements involved in ToM, such as the type of information used to make an inference, the depth and detail making up the inference itself, and the attribution of and response to uncertainty in a social interaction. Overall, the individual differences observed here are consistent with the notion of ad hoc ToM and use of general mental models as proposed in earlier chapters: Typically, inferences were made spontaneously and on demand.

4.3.8 Modelling Perspectives

From a modelling perspective, these theoretical propositions translate into implications for the mechanisms involved in ToM. A helpful heuristic to model ToM is the re-use of mental representations from similar previous encounters or more generally applicable situations rather than making new mental models from scratch each time they are required. Selective ToM use by situational demands and at different levels of detail may save cognitive costs and simultaneously cause the emergence of variety in observed ToM patterns.

The above themes offer more specific insights into what different ToM strategies may cause the observed variety in ToM manifestations. However, it is not clear what mechanisms deter-

mine on which basis one strategy will be used over another. For example, one person might reason about the other's strategy, another about the other's rewards and motivations, and a third about patterns in observed behaviour. These differences may be random, inherent differences in preferences and/or abilities, or differences based on previous learning. Importantly, however, the present data suggests that there is not one consistent process that makes up ToM, but rather there are a variety of different ways in which reasoning about others' minds (Hughes et al., 2005) can manifest. It may be intuitive to suggest investigating the cause of the different observed ToM manifestations by implementing a model that uses those different strategies and studying the different performances in comparison with human ToM patterns. This is beyond the scope of the present work but planned to be investigated in a future study.

Finally, this study indicates fragmented connections between the different elements involved in ToM. In the larger context of interconnected processes jointly making up ToM, the data points to partial execution of ToM, at least at different moments in time. With the overall ToM process involving observations, representations and maintenance of beliefs, and predictions of another's mental states and behaviours, the results indicate that different ToM elements may not all be used together fluently in any given moment, which should be considered in the implementation of a ToM model.

4.4 Insights

This study primarily offers insights about the large variance of ToM manifestation, different sources of ToM content, and ToM in practice.

Recall the questions specific to this chapter:

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

- Humans are complex and there are plenty of possible inferences to be made about others' mental states. What cognitive dynamics and heuristics affect the determination of specific inferences?
- What cognitive dynamics and heuristics characterise ToM variety?
- What cognitive dynamics and heuristics contribute to maximising ToM efficiency in social interactions?

The results offer the following suggestions answer these questions:

Theory of Mind Heuristics in Interactions

- ToM is fragmented, not necessarily performed in depth, and highly varied across individuals and situations
- Mental model content is retrieved from many different sources
- ToM reasoning does not directly translate into decisions and actions
  - There is often a discrepancy between cognitive representations and behaviour
  - Even with sparse information, certainty and confidence are dominant characteristics of applied ToM

Like previous studies, these insights are propositions of heuristics characterising ToM use in different elements underlying the overall process. This chapter outlines heuristics associated with the formation of mental models with different types of information and the effect of mental models on decisions and actions (highlighted in Figure 4.15).

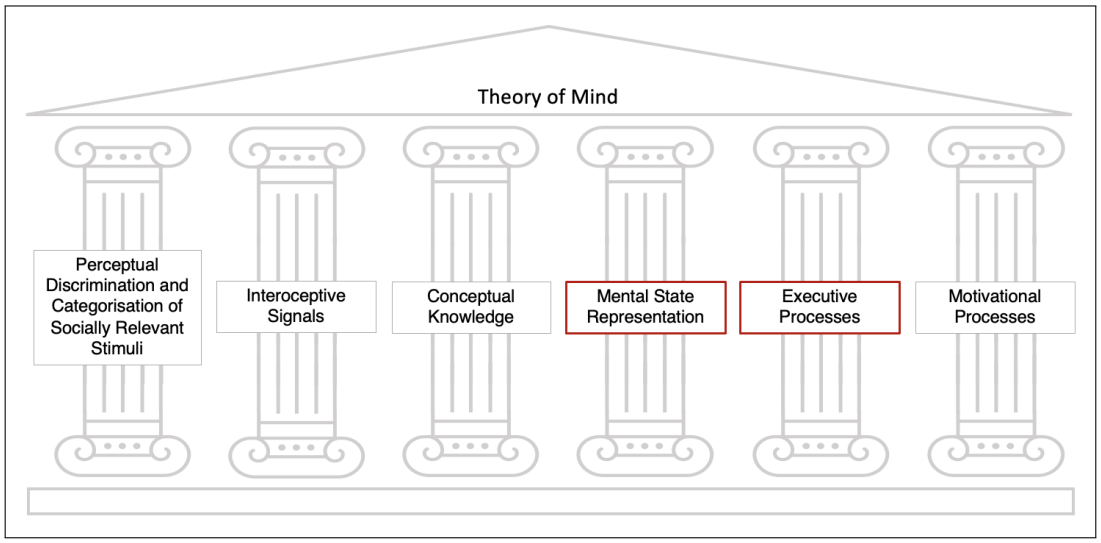


Figure 4.15: ToM elements considered in this thesis, based on Schaafsma et al. (2015).

Firstly, the findings of this study propose ToM processes to commonly be incomplete. Furthermore, the qualitative results suggest that in a simple strategic game there was not a lot of depth and recursion in participants' ToM inferences. In many instances no ToM-like reasoning was recorded at all, possibly because ToM inferences were too effortful for the benefit that accurate reasoning would have provided or the participants did not verbalise their inferences. From a perspective of bounded rationality, the human brain faces the challenge to invest available resources where and when this investment will be rewarded. More ToM is not always better, if it takes a lot of energy and provides little useful information. This work therefore suggests for the mental model component (in 4.15) to be conceptualised as functional even when repre-

sentations are few or incomplete. It is also proposed that human ToM involves a tendency to keep mental state representations as minimal as possible but can engage in more elaborate inferences when required and thereby create the variety in the extent of ToM on demand, as observed experimentally.

Secondly, the study at hand offers insights on what type of information can influence mental models and add to the individual differences observed in applied ToM. The following sources of information were identified and discussed: *Other's behaviour*, *Egocentrism*, *General mental model*, *Game logic*, and *Goals*. These are various aspects of either current experiences or beliefs stored in memory considered to different extents in the process of forming a mental model, depending on personal, social, or situational priorities. For example, the use of egocentric information or other already stored general mental models may speed up the formation a mental model but come with less accuracy than for example observing and considering direct observations of another person's behaviour. The hypothesis here is that depending on what is important or available for a given person in a given moment, the investment of cognitive energy into different aspects of ToM can differ substantially.

Thirdly, the study findings indicate discrepancies between the ToM included in verbal descriptions and subsequent behaviour. This indicates that rather than inferences directly affecting decisions and actions, there may still be other criteria affecting the decision such as risk. Alternatively, there may be a cost to the integration of a ToM belief into a subsequent decision calculation. Such a partial execution of ToM can reduce cognitive energy to the more essential elements and on demand, rather than continuously engaging in all ToM components.

To summarise, ToM differs across individuals and situations in various ways. This variance appears to occur at different stages, such as the initial attention to and perception of stimuli, the information considered, at the stage of inference calculation itself, or the extent to which mental models affect behaviour. At the stage of a person's current experience, different people will notice and perceive very different things in the same given situation, both internally and externally. Different individuals will have different beliefs about the world, which may again be applied to a situation in various ways. The representation of another person can then be affected by those elements in different ways. Different sources of information can be considered to various extents in this process of shaping mental models. Finally, the results at hand suggest that there are different degrees to which mental models can affect decisions and actions, causing potential discrepancies between what a person thinks, says, and does. Rather than reflecting an objective reality, ToM processes appear to be highly affected by an individual's own subjective experience and biased memories. They may be shaped to meet a person's own unique needs and priorities in navigating the social world, even in a simple competitive game.

Chapter 5

General Discussion

This chapter will summarise the insights discovered in previous chapters of this thesis. These insights will be discussed with regards to both the wider body of ToM literature and the conceptual ToM framework guiding this work.

5.1 Summary of Insights

The three experimental chapters of this thesis have contributed to the research of ToM by providing new perspectives and suggestions with regards to 1) inferential underpinnings of ToM, 2) the role of stereotypes in ToM efficiency, and 3) heuristics shaping the use of ToM in interactions.

Heuristics in Mental Model Manifestation

- Mental models can be generated on an ad hoc basis
- Different mental states can be inferred from different features of observed behaviour
  - Behaviour x → outcome x → goal x
  - Behaviour x → outcome x → plan x → intention x
  - Behaviour x → x is true → belief x
  - Regular behaviour x → habit x
  - Behaviour x → subjective evaluation x → opinion x
  - Behaviour x → choice of x over y → preference for x
- There are general principles that guide the determination of mental state inferences
  - People seek out what they like and avoid what they dislike
  - Behaviour is rooted in an underlying set of values and motivations
  - A person who is similar or different in one way is similar or different in other ways
  - Mental states align with each other and with the person's behaviour

The first study presented in this thesis explores verbal responses about mental states more generally. Findings suggest that mental state representations, especially about strangers, may be generated on an ad hoc basis. Furthermore, specific inferential patterns are proposed that describe what aspects of observed behaviour may contribute to what component of mental state representations, such as others' goals, beliefs, or preferences. Finally, the study suggests that this mapping of more specific observations and possible inferences could be guided by more general principles guiding a person's understanding of human behaviour, namely 1) People seek out what they like and avoid what they dislike, 2) Behaviour is rooted in an underlying set of values and motivations, 3) A person who is similar or different in one way is similar or different in other ways, and 4) Mental states align with each other and with the person's behaviour. These principles are here conceptualised as heuristics narrowing down the possibilities of what mental state may underlie a behaviour.

Stereotypes as Heuristics in Theory of Mind

- Mental model content is tied to social costs (group membership) and preparedness (predicting environmental threats)
- Informing mental models with existing and generalised mappings is less cognitively costly
  - Beliefs can be re-used rather than having to be newly formed
  - Stereotype-based mental models are generalised and therefore simpler to process
- The demands to change existing belief structures and mental models need to outweigh the required cognitive effort
  - Inconsistency across mental models is uncomfortable
  - Cognitive restructuring requires cognitive energy
  - Breaking down stereotypes involves representation of complex individual experiences

The second study supporting this thesis was concerned with stereotypes as a heuristic to reduce ToM costs by re-using information and simplifying representations of others. The study shows the role of social group membership and preparedness as factors that influence the way in which a person represents others. Furthermore, the study illustrated the effort and energy required to change existing mental models, particularly with changes from simper to more complex and individual mental models. This effort is typically rewarded with a personal reward that is deemed valuable enough to invest the resources necessary for the cognitive changes.

Theory of Mind Heuristics in Interactions

- ToM is commonly underused
- A way to reduce cognitive costs is using a model of the self to inform models of others
- ToM is fragmented, not necessarily performed in depth, and highly varied across individuals and situations
- Mental model content is retrieved from many different sources
- ToM reasoning does not directly translate into decisions and actions
  - There is often a discrepancy between cognitive representations and behaviour
  - Even with sparse information, certainty and confidence are dominant characteristics of applied ToM

The third and fourth study making up this thesis are concerned with other heuristics that contribute to more efficient ToM in interactions. Study 3 provides evidence for minimal and underused ToM and illustrates the phenomenon of egocentrism (applying the mental model of oneself to make sense of others) as a way of reducing the resources required to perform ToM. Study 4 suggests the possibility of little ToM use in the interactive strategic game played in the study. It also proposes that reasoning about others' intentions or actions may not necessarily translate into decisions and actions. Furthermore, the study highlights large variety in ToM and identifies various sources of information that can be used to generate social inferences.

5.2 Theoretical Discussion

The first chapter of this thesis introduced a perspective of cognitive costs which was adopted in this work to explore the role and effect of heuristics in ToM processes. Beyond these processes requiring energy, the 2nd study (The Human Library) suggests social or personal costs attached to violating social norms or being unprepared, or uncertain. The four studies supporting this work identify various ways in which cognitive costs may be reduced at the various processing stages of ToM.

5.2.1 Costs and Heuristics

Attending to the environment and making observations about a stimulus or another person is cognitively costly. Here it is suggested that this cognitive cost can be reduced by considering already existing memories rather than paying attention to form new ones. This thesis has sug-

gested different heuristics that may be used in ToM to reduce the resources required to perform the ability. Various ways were identified that may reduce the cognitive costs required for mental model formation.

Recall the overall research question of this thesis:

Research Question:

What cognitive dynamics and heuristics characterise efficient yet robust ToM?

There appear to be overall tendencies towards simpler and minimal rather than elaborate mental models, or already existing rather than newly established representations, thus making inference less costly. Moreover, the use of egocentrism and generalisations were identified as heuristics reducing the cognitive costs spent on forming mental models by, for example, reducing the need for additional observations, mapping them into a coherent set of beliefs, and maintaining those beliefs over time.

Another element of ToM which heuristics were identified for is the element of conceptual knowledge to be established and/or re-structured in memory. This thesis suggests that stereotypes may be a mechanism that reduce costs spent at this stage by means of involving overall simpler and more generally applicable representations. Furthermore, it is suggested that the maintenance of coherence across beliefs, concepts, and representations makes it more straight forward to generate abstractions and generalisations across concepts and situations.

Moreover, ad hoc ToM is proposed as a mechanism reducing the cognitive costs required to form mental models. The ad hoc nature is thereby conceptualised in the sense that situational demands determine whether a mental state representation is formed in the first place, and if it is necessary, the relevant details can be retrieved from various sources in that moment, rather than having to be established and maintained in advance. Furthermore, this thesis suggests that the information making up a mental model can be retrieved from various sources, depending on what is more available in a given moment. This thesis also identified various inference patterns and general principles that were proposed to guide the formation of mental models in a heuristic manner to narrow down the alternatives included in the process.

Finally, executive processes were suggested to be characterised by selectivity from cognition to action, and motivational processes were proposed to be influenced by heuristics linked to social and personal costs, determining mental models and social evaluations.

Table 5.1 summarises these heuristics with regards to the different elements of ToM considered in this work, and indicates what studies in this thesis provide supporting evidence. The subsequent section relates these insights back to the issues raised in the introduction.

Cost-saving heuristic	Supporting study
Element: Perceptual Discrimination and Categorisation of Socially Relevant Stimuli	
Re-using already existing memories	Ch 3 - The Human Library
Element: Interoceptive Signals	
Egocentrism	Ch 4 - Cutting Corners in Theory of Mind
Element: Conceptual Knowledge	
Generalisations (specifically, Stereotypes)	Ch 3 - The Human Library
Maintaining coherence	Ch 3 - The Human Library
Element: Mental State Representation	
Ad Hoc ToM	Ch 2 - Ad Hoc Theory of Mind
Minimalist Theory of Mind	Ch 4 - Cutting Corners in Theory of Mind
Use of information from multiple sources to fill gaps	Ch 4 - Theory of Mind in a Strategic Game
Use of inference patterns	Ch 2 - Ad Hoc Theory of Mind
Element: Executive Processes	
ToM may affect decisions but not necessarily	Ch 4 - Theory of Mind in a Strategic Game
Element: Motivational Processes	
Social Motivation	Ch 3 - The Human Library
Cost of Change	Ch 3 - The Human Library

Table 5.1: Heuristics associated with different processes and the respective supporting studies.

5.2.2 Modelling Perspectives

Bayesian approaches have shown very promising performances in modelling ToM (Baker et al., 2011). However, using these probabilistic models may require a lot of resources. The process of creating them can be costly in terms of observations required and computation of probabilities. Further, the process of generating inferences can also be computationally costly, especially for models with a large number of variables. For humans, this process would be intractable (Kwisthout and Van Rooij, 2020). Importantly, the richness of variables affecting ToM manifestations as discussed in this work are very difficult to reconcile with a modelling approach that requires exact specifications of all variables involved, such as an encapsulated Bayesian model. Instead, human minds are more selective from the start: Even though they have fundamentally predictive characteristics (Clark, 2015), they also employ more explicit heuristics. This

accounts for the limit to available cognitive resources (Lieder and Griffiths, 2020) by reducing the cognitive computations required for ToM.

These heuristics are explored throughout this thesis and what was found here is summarised in the section above. These patterns were identified as mechanisms that reduce cognitive costs associated with ToM. Importantly, this speaks for their efficiency but does not mean that they are always ideal. Different characteristics of mental representations can be prioritised. For example, the accuracy of a mental model may be sacrificed to maintain coherence with other mental models or accommodate social norms. It may be helpful to relate this back to the Barista example given at the end of the introduction: The customer could have theoretically identified the true reason for why the barista asked her for her name. However, she focussed on a subset of the information available to her and came to a false conclusion. She prioritised the quick response regarding her relationship status with strangers in public over correctly identifying coffee-related intentions. From a standpoint of accuracy, this is not ideal. However, the concept of bounded rationality suggests that a person's own individual values and goals determine the way in which available resources are spent optimally.

Another aspect of ToM touched on in this thesis is recursion. It is a fundamental component of social inference models such as PsychSim (Marsella et al., 2004; Pynadath and Marsella, 2005), reasoning about other's beliefs at various levels of recursive depth. This, however, would also be impossible for humans to do, given the complexity of social situations and interactions. Gmytrasiewicz et al. 1991 suggest that beyond a certain level of depth, recursion only minimally improves a model's ToM performance. In line with this conclusion, this thesis suggests that humans may not always reason very deeply about others in practice. Even though it is theoretically possible to reason about others in depth and it seems like a fundamental aspect of what makes up ToM reasoning, this thesis proposes that most of practical ToM is based on heuristic shortcuts, generalisations, and biases, rather than deep, elaborate, and complete ToM, as it is theoretically conceptualised.

Finally, as outlined in the introduction, Pöppel and Kopp (2019) have developed a model including a mechanism that switches between simpler and more complex mental models. Their approach postulates that humans switch to simple mental models whenever possible, with room for more complex representation when required. The tendency towards simpler mental model is also reflected in the findings of this thesis, for example from the Human Library study. However, the present work suggests that switching between different models is not the whole story. Rather, this thesis proposes that sometimes models are only created on an ad hoc basis or even not formed at all. It also elaborates more on the wider ToM processes, such as the retrieval of information from different sources, the type of existing general social biases, and the dynamics with other beliefs memories that shape a person's contextualised and coherent experience.

The insights discussed in this thesis are based on rules and mechanisms. It would be interesting to combine these insights with a network-like architecture to account for the biological

structure of the human brain (like for example Blouw et al, 2016). Hopefully the mechanistic understandings at hand can help inform subsequent models based on more network-like architectures. Many of the discussed propositions align with each other, for example egocentrism and ad hoc ToM, and may eventually lead to the establishment of more general principles that can be realised in a network-based model. Here, however, it has been very helpful to focus on understanding the mechanisms and processes that underlie the behaviours and patterns that has been observed in ToM research.

The summary and discussion of the insights emerging from this thesis poses a general question for AI: Are these heuristics and biases the way forward in modelling ToM? A perfect implementation of human ToM in a model would certainly reduce the costs required for the processes involved. However, human ToM appears to be characterised by fundamental errors and biases, such as egocentrism or stereotypes. The barista scenario from the beginning exemplifies how easily human ToM can go wrong and often this remains undetected on a daily basis. For example, researchers have suggested that a proportion of mental health issues is rooted in false attribution of others' intentions (Zainal and Newman, 2018). Furthermore, wrongly attributing another's mental states may cause misunderstandings and conflict (Chambers and De Dreu, 2014). The human library study illustrates the difficulty in changing mental models of others and potential consequences of misattribution.

It may certainly be beneficial to develop AI that can conceptualise and process that humans show these behavioural patterns and biases, but it may be counterproductive to develop algorithms that produce the same errors and biases, such as stigma or discrimination. The involvement of different elements also depends on the function of the respective AI, determining which algorithms should and should not be implemented. For example, for an AI that interacts with humans in a low-stakes context, it may be beneficial to implement minimal ToM for human likeness and include general inference patterns and an ad hoc architecture (study 1) to reduce processing costs and constrain the processes required. At the same time, this implementation may benefit from avoiding phenomena such as stereotypes and social biases, to reduce the risk of facilitating or even creating discrimination and stigma. Applications in psychological research, on the other hand, such as social simulations to understand human behavioural patterns, may benefit from human-like biases. This would include stereotypes and fragmented ToM, to simulate human behaviour as accurately as possible. In a high-stakes context, however, accurate prediction of human behaviour can be more important than saving resources and reflecting human-like ToM processes. In those instances, it may be beneficial to model human ToM heuristics such as reduced ToM, re-use of memories, egocentrism, stereotypes, and tendencies for coherence with more elaborate and comprehensive modelling strategies, and invest the required resources to achieve accurate results.

Overall, these modelling perspectives may be elaborated in future work by identifying algorithms that reflect the heuristics identified here. For example, concretely implementing ad

hoc notions as a heuristic would involve the development of demand-based architectures and the identification of inference patterns that fit into such an architecture. This next step of finding concrete grounds for the heuristics proposed in this work would have several benefits: Firstly, the heuristics hypothesised here could be tested and compared to human behaviour. Secondly, different scenarios could be simulated to investigated more concretely when the computational use of which heuristics may be helpful and when not. Finally, a working model of the different heuristics at hand may be helpful for the future study of ToM processes, refining theories, and generating new hypotheses about other heuristics and cognitive patterns.

5.2.3 Psychological Perspectives

This thesis provides insights into the general theoretical understanding of ToM by means of exploring underlying components and mechanisms. It suggests that many different cognitive processes are involved in shaping ToM manifestations in practice. Particularly, the focus of this work is on the mechanisms that allow humans to reason about others in the rich and complex social spaces they enter every day in a sustainable and resource-optimal way. This thesis proposes that ToM is highly context-dependent and situated: a person's goals and situational demands shape both the extent to which they can reason about others and what direction this reasoning process will take. Multiple elements have been identified and discussed in this work that were proposed to act as heuristics to allow this situatedness and let a person flexibly make sense of others' behaviour in the context of their own needs, resources, and experience of the world.

These different elements have likely evolved over a long time and contribute to the large variety observed in human ToM today. It is more and more accepted by ToM researchers that neither Theory Theory nor Simulation Theory can explain ToM results on their own and a more pluralist approach (Gallagher and Fiebich, 2019) accounts better for a variety of phenomena. Similarly, this thesis proposes that various mechanisms and heuristics contribute to the complexity and robustness of ToM. Importantly, this thesis has explored these heuristics, within a conceptual ToM framework to theoretically inform the body of model-focussed research on the topic.

The ToM heuristics proposed and discussed here are mechanisms that allow an individual to reason quickly and smoothly in a complex social environment with the cognitive resources they have available and the demands they are facing in a given situation. Thereby, they may prioritise different opportunity costs, and creating very different manifestations of ToM. For example, egocentric reasoning about another's beliefs may lead to inaccuracy and errors but provide a sense of certainty in navigating the own social space. From a perspective of bounded rationality (Lieder and Griffiths, 2020), even ToM failures may be optimal for a person when deep and elaborate ToM requires more resources than it creates rewards.

The current perspective on ToM can be related back to other ToM issues discussed in previous literature, such as the dual process account (Chaiken and Trope, 1999; Kahneman, 2011).

The present conceptual framework does not highlight the same clear-cut distinction between slow and fast mechanisms as other work has but it does not stand in contrast to the approach. Biases such as egocentrism or stereotypes may certainly fall into the category of fast ToM whereas more deliberate and careful consideration of the environment and beliefs that may be less intuitive or accessible may be classified as slow ToM. However, the focus of the present work is not to distinguish between different types of ToM processes as such, but rather identify the range of different strategies humans can employ to make sense of their social environment in an efficient and adaptable manner. For example, consider the barista example: The customer's immediate reaction to the barista's question was a mistaken interpretation that was most available to her in that moment. When she was corrected, she re-assessed the situation with the new information more carefully. It is possible that the processes in this example can be characterised into the different elements of the dual process perspective, but that is not the focus here. What this work aims to illustrate is how many different factors affect ToM and how versatile the brain's strategies are that let an individual account for available resources and needs.

One of the discussed heuristics is egocentrism, which has been studied previously for example by Keysar et al. (2003). Interestingly, Southgate (2020) proposed that very young infants are fully altercentric, considering everything from other people's perspectives. From the perspective of the framework and heuristics this thesis discusses, this makes sense because they have not learned and experienced the same information that makes up the beliefs and biases that are available to an individual later-on in life. However, from a bounded rationality perspective, the question would still remain what in the vast space of possibilities of others' mental space guides an infant in generating their respective mental models. The development of egocentrism as a ToM heuristic, like it was demonstrated in the third study supporting this work, may constrain that selection. It may reduce the cognitive demands that increase as a child gets older and navigates more complex social stimuli. In contrast, Burge (2018) argues that there are other powerful explanations such as associative learning or reinforcement that can explain ToM-like behaviour. However, these possibilities are not necessarily exclusive. As the last study of this thesis suggests, it is very possible that the brain flexibly and momentarily adopts the more useful strategy out of all of the above.

Beyond ToM theory, the points of discussion in this work align with the study of more general cognitive mechanisms. Here, for example, abstraction and generalisation processes are proposed as efficiency mechanisms in ToM. By creating this link with basic organisations of cognitive concepts, this thesis argues that research on coherence (Thagard, 2002) or the organisation of beliefs and concepts (Barsalou, 2003), including ad hoc categories (Barsalou, 1983), is fundamentally connected with a person's social experience. Similarly, it is suggested here that principles such as assimilation and accommodation and their role for learning (Piaget, 1976) are relevant to with social reasoning processes. Overall, it suggests an angle of inter-connectivity across cognitive phenomena: This work proposes that human cognition has evolved to spend re-

sources optimally and gain resource advantages by linking different cognitive processes together to save cognitive costs where possible. This is in line with previous work on cognitive biases and heuristics (Kahneman, 2011; Clark, 2015). This inter-connective perspective also resonates with a situated action approach to cognition (Barsalou, 2020), which suggests interactions between brain, body, and the environment. Thereby, the human experience is very strongly shaped by situational factors, both internal and external. Similarly, Badcock et al. (2019) proposes that experiences are not isolated but rather contextualised and constrained by the boundaries of our senses.

Furthermore, the Human Library study illustrates the applicability of ToM research and insights to wider reaching issues, such as understanding stereotypes and other categorisation mechanisms. Applying the findings of the present work to these areas may aid the development of practical interventions in response to wider social issues, discrimination, and stigma. Particularly, understanding ToM as a phenomenon characterised by heuristics which may save cognitive costs but also come with reduced accuracy in understanding others, poses the question of whether this saving of costs is worth it and how humans can navigate the trade-off. The heuristics proposed in this work all have the advantage of saving costs and letting humans do the complex ability of reasoning about complex social scenarios very quickly and effortlessly. However, heuristic-based evaluations and attributions also appear to be less reliable and more prone to error. For example, stereotype-based thinking may provide structure and expectations in situation where not a lot of information is available, and therefore be perceived as comfortable by the individual. On the other hand, it may also create discrimination of whole social groups, and lead to misunderstandings or conflict. An interesting next step would be to identify what exact advantages and disadvantages characterise the practical use of the different heuristics proposed here. Developing a concrete guide to the advantages and disadvantages of different heuristics may be a worthwhile piece of future research to help individuals navigate the benefits and dangers of ToM heuristics and identify which may be helpful and which may not be.

It is essential to keep in mind that the term ToM as it is used in this thesis is based on one particular definition of the term, i.e. the explicit and cognitive generation of social inferences to reason about other's mental states and behaviour. The implications of this work are therefore also only concerned with this understanding of the concept and may not apply if ToM is defined differently to begin with. With this definition of the term, *Robust, Efficient, Dynamic Theory of Mind* has been outlined as an individualised and situation-based cognitive phenomenon, shaped by various structures and heuristics that facilitate social reasoning in consideration of a person's own goals, resources, and situational demands.

5.3 Conclusions and Future Directions

This thesis has focussed on conceptual considerations of ToM mechanisms to form a solid basis, a starting point, for the modelling of Robust, Efficient, Dynamic Theory of Mind. The previous chapters present a range of different heuristics proposed to characterise various elements that were suggested to be involved in ToM by Schaafsma et al. (2015).

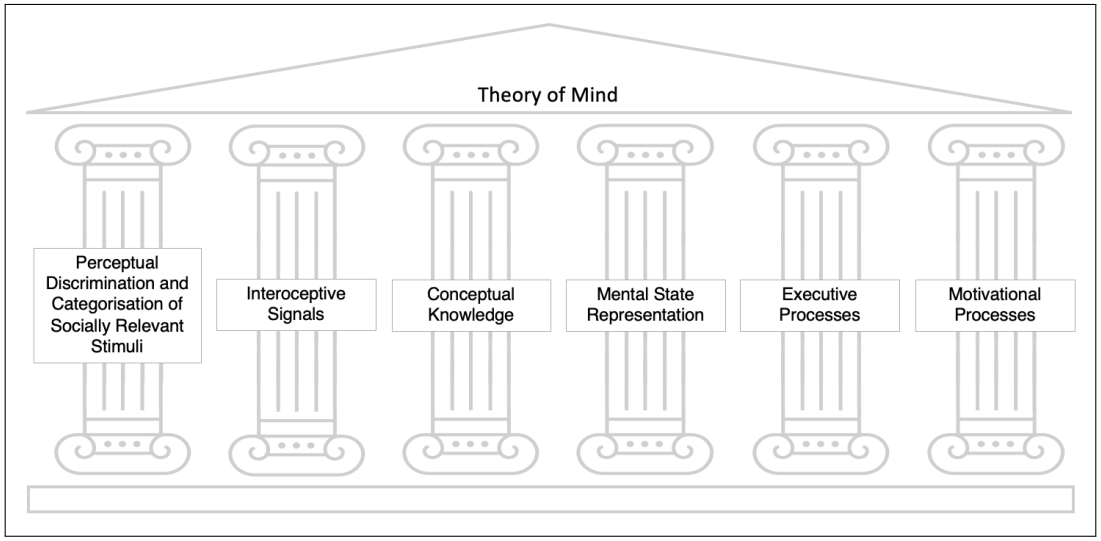


Figure 5.1: ToM elements, based on Schaafsma et al. (2015).

Most importantly, this work has offered insights about the complexity of ToM as a human ability to navigate social interactions. It has shown that there is no one way to do ToM and the very same characteristics that let humans use it so easily and quickly are the features that easily lead to mistakes and difficulty in social situations. Specifying the heuristics above and their place in the overall maze of ToM is an important first step towards computational modelling of human-like ToM.

Future research is planned to address the concrete implementation of such a model. Specifically, future work will aim to include computations that represent the heuristics presented here. For example, ad hoc ToM represents the notion that an individual does not even necessarily have a model of the person they interact with, only when this model is helpful based on momentary demands on the individual. The representation of another's mental states would therefore only be generated on demand as well. This representation can be see as a product of motivated inference (Kunda, 1990). Furthermore, it is proposed here that relevant information is pulled in from different sources, such as the current experience, previous beliefs about a similar instance, or egocentric information about oneself. These structures and pathways need to be established computationally. Search algorithms need to be selective, sensitive to goals and demands, distinguish between different sources of information, and retrieve relevant elements on a situational basis. It is planned to model the different heuristics explored in this work (Table 6.1), both in-

dividually and in combination, to compare and contrast model behaviour with human behaviour and assess performance of different models with different heuristic strategies.

In contrast to previous models of ToM, this work argues for an approach that is not fixed and encapsulated. Here it is proposed that the model an agent has of another might more appropriately be pulled together based on situational demands and available resources, rather than a discrete, distinct collection of representations of the other's presumed mental states. Furthermore, these elements may be retrieved from various sources and different approaches may be employed to perform the required inferences. Therefore, future work will be directed towards the development of algorithms and models that can offer the required flexibility to accommodate this ad hoc notion. The goal will be to study and explore how the complexity and various dimensions of the discovered theoretical underpinnings of ToM can be incorporated in a model, rather than adjusting it to fit in a fixed and encapsulated approach.

When a model is implemented successfully, this can then be employed to model more applied social issues and their complexity, such as stereotypes and stigma, and ideally lead to new strategies in the development of intervention strategies. It can also be used to include more accurate models of humans in AI more generally, which include representations of the biases and heuristics that humans typically engage in, such as stereotypical thinking, egocentric bias, or, situationally, no ToM altogether. An awareness in AI of the strengths and weaknesses in human cognition and their effects on social interactions may inform the improvement of human-AI interaction, especially in human-AI teamwork. This research also helps consider models of ToM from a more psychologically informed perspective. It sheds light on the psychological underpinnings of ToM and may help explain why some models do not reflect human-like ToM.

This thesis has established an account of ToM as characterised by variance and richness, including connections with many other cognitive processes. This perspective may encourage a development of novel and improved paradigms to study ToM. As Schaafsma et al. (2015) argue, the body of ToM is characterised by many different definitions of the concept and a variety of methods to test it. The complexity of factors shaping ToM manifestation established in the present work may contribute to the deconstruction and reconstruction of ToM that Schaafsma et al. (2015) call for. Particularly in line with the challenges in clearly recording instances of ToM that was noticed in this work, further study of precise measures that can more certainly confirm or rule out the use of ToM in social interactions is encouraged.

Finally, most of the research supporting this thesis is exploratory. It has contributed to the body of literature on ToM by providing many new theoretical considerations of what heuristic phenomena may characterise ToM in practice. At the same time, there is now a need for more confirmatory research assessing the suggestions presented in this thesis and testing the theories and perspectives that have emerged. It will be fruitful in future studies to employ quantitative methods to test the primarily qualitative results illustrated here. Many of the explored mechanics would need further quantitative testing to establish their relationships. Causality and

independence would have to be investigated, otherwise implementation in any model would be extremely difficult. This would be a data-intensive endeavour that would require many more years of research than is allocated to a single PhD thesis.

Specific objectives for future research may therefore be as follows:

- Development of algorithms reflecting the heuristics identified in this work.
- Using a computational model based on these heuristics to test their usefulness and applicability in ToM research.
- Investigating how different ToM heuristics can be realised in combination.
- Identification of advantages and disadvantages of the different ToM heuristics in practice, towards developing education materials.
- Development of a concrete measure of ToM use.

Engaging and committing to this work over the course of several years has been both challenging and truly intriguing. It work has raised and tackled very specific questions about the execution of ToM in practice and at the same time required reflection on very broad issues of human nature. I hope that any future research I conduct will spark the same engagement and passion, and facilitate my own development as a critical thinker and a social individual.

Bibliography

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

Alfonso, B., Pynadath, D. V., Lhommet, M., and Marsella, S. (2015). Emotional Perception for Updating Agents' Beliefs. *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 201–207.

Allport, G. W., Clark, K., and Pettigrew, T. (1954). The nature of prejudice. Publisher: Addison-wesley Reading, MA.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, 65(5):825–839. Publisher: SAGE Publications Sage UK: London, England.

Apperly, I. A. and Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4):953–970.

Avramides, A. (2019). Perception, Reliability, and Other Minds. In *Knowing other minds*, pages 107–126. Oxford University Press.

Avramides, A. and Parrott, M. (2019). *Knowing other Minds*. Oxford University Press.

Badcock, P. B., Friston, K. J., and Ramstead, M. J. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of life Reviews*, 31:104–121. Publisher: Elsevier.

Baker, C. L., Saxe, R. R., and Tenenbaum, J. B. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33(33):2469–2474.

Banich, M. T. and Caccamise, D. (2011). *Generalization of knowledge: Multidisciplinary perspectives*. Psychology Press.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11. Issue: 1.

Bardi, L., Desmet, C., Nijhof, A., Wiersema, J. R., and Brass, M. (2016). Brain activation for spontaneous and explicit false belief tasks overlaps: new fMRI evidence on belief processing and violation of expectation. *Social Cognitive and Affective Neuroscience*, page nsw143.

Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International review of research in mental retardation*, 23:169–184. Publisher: Elsevier.

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21:37–46.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, 11(3):211–227. Publisher: Springer.

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1435):1177–1187. Publisher: The Royal Society.

Barsalou, L. W. (2019). Establishing generalizable mechanisms. *Psychological Inquiry*, 30(4):220–230. Publisher: Taylor & Francis.

Barsalou, L. W. (2020). Categories at the interface of cognition and action. In *Building Categories in Interaction: Linguistic Resources at Work*. John Benjamins, Amsterdam.

Bartlett, T. (2017). Can we really measure implicit bias? Maybe not. *The chronicle of higher education*, 63(21):B6–B7.

Bianco, F. (2022). *Theory of mind across biological and artificial embodiment: theory, experiments and computational models*. PhD Thesis, University of Essex.

Blouw, P., Solodkin, E., Thagard, P., and Eliasmith, C. (2016). Concepts as semantic pointers: A framework and computational model. *Cognitive science*, 40(5):1128–1162. Publisher: Wiley Online Library.

Blum, C., Winfield, A. F., and Hafner, V. V. (2018). Simulation-based internal models for safer robots. *Frontiers in Robotics and AI*, 4:74. Publisher: Frontiers.

Bolstad, C. A. and Endsley, M. R. (1999). Shared mental models and shared displays: An empirical evaluation of team performance. In *proceedings of the human factors and ergonomics society annual meeting*, volume 43, pages 213–217. SAGE Publications Sage CA: Los Angeles, CA. Issue: 3.

Bratman, M. (1987). *Intention, plans, and practical reason*, volume 10. Harvard University Press, Cambridge, MA.

Burge, T. (2018). Do infants and nonhuman animals attribute mental states? *Psychological Review*, 125(3):409–434.

Caligiore, D., Tommasino, P., Sperati, V., and Baldassarre, G. (2014). Modular and hierarchical brain organization to understand assimilation, accommodation and their relation to autism in reaching tasks: a developmental robotics hypothesis. *Adaptive Behavior*, 22(5):304–329. Publisher: Sage Publications Sage UK: London, England.

Carruthers, P. and Smith, P. K. (1996). *Theories of theories of mind*. Cambridge University Press.

Chaiken, S. and Trope, Y. (1999). *Dual-process theories in social psychology*. Guilford Press.

Chambers, J. R. and De Dreu, C. K. (2014). Egocentrism drives misunderstanding in conflict and negotiation. *Journal of Experimental Social Psychology*, 51:15–26. Publisher: Elsevier.

Charters, E. (2003). The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal*, 12(2).

Choi, Y.-j., Mou, Y., and Luo, Y. (2018). How do 3-month-old infants attribute preferences to a human agent? *Journal of Experimental Child Psychology*, 172:96–106.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *BEHAVIORAL AND BRAIN SCIENCES*, 36(3):1–73.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Conway, J. R. and Bird, G. (2018). Conceptualizing degrees of theory of mind. *Proceedings of the National Academy of Sciences*, 115(7):1408–1410.

Conway, J. R., Catmur, C., and Bird, G. (2019). Understanding individual differences in theory of mind via representation of minds, not mental states. *Psychonomic Bulletin & Review*, 26(3):798–812.

Coull, J. T. (2004). fMRI studies of temporal attention: allocating attention within, or towards, time. *Cognitive Brain Research*, 21(2):216–226. Publisher: Elsevier.

Csibra, G. (2017). Cognitive science: Modelling theory of mind. *Nature Human Behaviour*, 1(4):0066.

Deschrijver, E., Bardi, L., Wiersema, J. R., and Brass, M. (2016). Behavioral measures of implicit theory of mind in adults with high functioning autism. *Cognitive Neuroscience*, 7(1-4):192–202.

Driver, J. (2001). A selective review of selective attention research from the past century. *British journal of psychology*, 92(1):53–78. Publisher: Wiley Online Library.

Duehr, E. E. and Bono, J. E. (2006). Men, women, and managers: are stereotypes finally changing? *Personnel psychology*, 59(4):815–846. Publisher: Wiley Online Library.

Dunn, T. L., Inzlicht, M., and Risko, E. F. (2019). Anticipating cognitive effort: roles of perceived error-likelihood and time demands. *Psychological research*, 83:1033–1056. Publisher: Springer.

Edwards, J. A. and Weary, G. (1998). Antecedents of causal uncertainty and perceived control: A prospective study. *European Journal of Personality*, 12(2):135–148. Publisher: Wiley Online Library.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and Rasmussen, D. (2012). A large-scale model of the functioning brain. *science*, 338(6111):1202–1205. Publisher: American Association for the Advancement of Science.

Elliot, A. J. and Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of personality and social psychology*, 67(3):382. Publisher: American Psychological Association.

Fechner, H. B., Schooler, L. J., and Pachur, T. (2018). Cognitive costs of decision-making strategies: A resource demand decomposition analysis with a cognitive architecture. *Cognition*, 170:102–122. Publisher: Elsevier.

Festinger, L. (1957). *A theory of cognitive dissonance*, volume 2. Stanford university press.

Fiebich, A. and Coltheart, M. (2015). Various ways to understand other minds: Towards a pluralistic approach to the explanation of social understanding. *Mind & Language*, 30(3):235–258. Publisher: Wiley Online Library.

Freundlieb, M., Kovács, M., and Sebanz, N. (2015). When do humans spontaneously adopt another's visuospatial perspective? *Journal of Experimental Psychology: Human Perception and Performance*, 42(3):401–412.

Frith, C. (2013). *Making up the mind: How the brain creates our mental world*. John Wiley & Sons.

Frith, U. and Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):459–473.

Frith, U. and Happé, F. (1994). Autism: beyond "theory of mind". *Cognition*, 50(1-3):115–132.

Gallagher, S. and Fiebich, A. (2019). Being pluralist about understanding others: Contexts and communicative practices. In *Knowing other minds*, pages 63–78. Oxford University Press.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501. Publisher: Elsevier.

Gandhi, K., Fränken, J.-P., Gerstenberg, T., and Goodman, N. D. (2023). Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448*.

Garbarino, E. C. and Edell, J. A. (1997). Cognitive effort, affect, and choice. *Journal of consumer research*, 24(2):147–158. Publisher: The University of Chicago Press.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., and others (2011). Differences between tight and loose cultures: A 33-nation study. *science*, 332(6033):1100–1104. Publisher: American Association for the Advancement of Science.

Giesler, M. A. (2022). Humanizing Oppression: The Value of the Human Library Experience in Social Work Education. *Journal of Social Work Education*, 58(2):390–402. Publisher: Taylor & Francis.

Gmytrasiewicz, P. J. and Doshi, P. (2004). Interactive pomdps: Properties and preliminary results. In *International Conference on Autonomous Agents: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-*, volume 3, pages 1374–1375.

Gmytrasiewicz, P. J. and Durfee, E. H. (1995). A Rigorous, Operational Formalization of Recursive Modeling. In *ICMAS*, pages 125–132.

Gmytrasiewicz, P. J., Durfee, E. H., and Wehe, D. K. (1991). A Decision-Theoretic Approach to Coordinating Multiagent Interactions. *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, 91:62–68.

Goodie, A. S., Doshi, P., and Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, 25(1):95–108. Publisher: Wiley Online Library.

Greggor, A. L., Thornton, A., and Clayton, N. S. (2017). Harnessing learning biases is essential for applying social learning in conservation. *Behavioral Ecology and Sociobiology*, 71:1–12. Publisher: Springer.

Groyecka, A., Witkowska, M., Wróbel, M., Klamut, O., and Skrodzka, M. (2019). Challenge your stereotypes! Human Library and its impact on prejudice in Poland. *Journal of Community & Applied Social Psychology*, 29(4):311–322. Publisher: Wiley Online Library.

Guzman, A. L. and Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New media & society*, 22(1):70–86. Publisher: Sage Publications Sage UK: London, England.

Gweon, H., Young, L., and Saxe, R. R. (2011). Theory of Mind for you, and for me: behavioral and neural similarities and differences in thinking about beliefs of the self and other. In *Proceedings of the annual meeting of the cognitive science society*, volume 33. Issue: 33.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., and Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *science*, 317(5843):1360–1366. Publisher: American Association for the Advancement of Science.

Herzig, A., Lorini, E., Perrussel, L., and Xiao, Z. (2017). BDI Logics for BDI Architectures: Old Problems, New Perspectives. *KI - Künstliche Intelligenz*, 31(1):73–83.

Heyes, C. M. and Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, 344(6190):1243091–1243091.

Hill, M. E. and Augoustinos, M. (2001). Stereotype change and prejudice reduction: short-and long-term evaluation of a cross–cultural awareness programme. *Journal of Community & Applied Social Psychology*, 11(4):243–262. Publisher: Wiley Online Library.

Hilton, J. L. and Von Hippel, W. (1996). Stereotypes. *Annual review of psychology*, 47(1):237–271. Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., and Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child development*, 76(2):356–370. Publisher: Wiley Online Library.

Jacob, P. (2019). Challenging the Two-systems Model of Mindreading. In *Knowing other minds*, pages 79–106. Oxford University Press, 1 edition.

Jameson, A. (1996). Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-Adapted Interaction*, 5:193–251. Publisher: Springer.

Jetten, J., Branscombe, N. R., Haslam, S. A., Haslam, C., Cruwys, T., Jones, J. M., Cui, L., Dingle, G., Liu, J., Murphy, S., and others (2015). Having a lot of a good thing: Multiple important group memberships as a source of self-esteem. *PloS one*, 10(5):e0124609. Publisher: Public Library of Science San Francisco, CA USA.

Jonker, C. M., Riemsdijk, M., and Vermeulen, B. (2010). Shared mental models. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 132–151. Springer.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kajić, I., Schröder, T., Stewart, T. C., and Thagard, P. (2019). The semantic pointer theory of emotion: Integrating physiology, appraisal, and construction. *Cognitive Systems Research*, 58:35–53.

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89(1):25–41.

Kiger, M. E. and Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical teacher*, 42(8):846–854. Publisher: Taylor & Francis.

Kobayashi, C., Glover, G. H., and Temple, E. (2007). Cultural and linguistic effects on neural bases of 'Theory of Mind' in American and Japanese children. *Brain Research*, 1164:95–107.

Koster-Hale, J. and Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, 79(5):836–848. Publisher: Elsevier.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3):480. Publisher: American Psychological Association.

Kurzban, R. and Neuberg, S. (2015). Managing ingroup and outgroup relationships. *The handbook of evolutionary psychology*, pages 653–675. Publisher: Wiley Online Library.

Kwisthout, J. and Van Rooij, I. (2020). Computational resource demands of a predictive Bayesian brain. *Computational Brain & Behavior*, 3:174–188. Publisher: Springer.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40. Publisher: Cambridge University Press.

Leppanen, J., Sedgewick, F., Treasure, J., and Tchanturia, K. (2018). Differences in the Theory of Mind profiles of patients with anorexia nervosa and individuals on the autism spectrum: A meta-analytic review. *Neuroscience & Biobehavioral Reviews*, 90:146–163. Publisher: Elsevier.

Leslie, A. M. (1991). The theory of mind impairment in autism: Evidence for a modular mechanism of development? In *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, pages 63–78. Basil Blackwell. Publisher: Basil Blackwell.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1. Publisher: Cambridge University Press.

Lim, B.-C. and Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 27(4):403–418. Publisher: Wiley Online Library.

Low, L.-F. and Purwaningrum, F. (2020). Negative stereotypes, fear and social distance: a systematic review of depictions of dementia in popular culture in the context of stigma. *BMC geriatrics*, 20(1):1–16. Publisher: BioMed Central.

Madva, A. and Brownstein, M. (2018). Stereotypes, prejudice, and the taxonomy of the implicit social mind1. *Noûs*, 52(3):611–644. Publisher: Wiley Online Library.

Marsella, S. C., Pynadath, D. V., and Read, S. J. (2004). PsychSim: Agent-based modeling of social interactions and influence. *Proceedings of the international conference on cognitive modelling*, 36:243–248.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., and Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2):273. Publisher: American Psychological Association.

Matsuka, T. and Corter, J. E. (2008). Observed attention allocation processes in category learning. *Quarterly Journal of Experimental Psychology*, 61(7):1067–1097. Publisher: SAGE Publications Sage UK: London, England.

Matz, D. C. and Wood, W. (2005). Cognitive dissonance in groups: the consequences of disagreement. *Journal of personality and social psychology*, 88(1):22. Publisher: American Psychological Association.

Medin, D. L. and Smith, E. E. (1984). Concepts and concept formation. *Annual review of psychology*, 35(1):113–138. Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Meltzoff, A. N. (2007). The 'like me'framework for recognizing and becoming an intentional agent. *Acta psychologica*, 124(1):26–43. Publisher: Elsevier.

Mullin, B.-A. and Hogg, M. A. (1999). Motivations for group membership: The role of subjective importance and uncertainty reduction. *Basic and applied social psychology*, 21(2):91–102. Publisher: Taylor & Francis.

Mulvey, K. L., Rizzo, M. T., and Killen, M. (2016). Challenging gender stereotypes: Theory of mind and peer group dynamics. *Developmental Science*, 19(6):999–1010. Publisher: Wiley Online Library.

Mushtaq, F., Bland, A. R., and Schaefer, A. (2011). Uncertainty and cognitive control. *Frontiers in psychology*, 2:249. Publisher: Frontiers Research Foundation.

Newen, A. and Wolf, J. (2020). The Situational Mental File Account of the False Belief Tasks: A New Solution of the Paradox of False Belief Understanding. *Review of Philosophy and Psychology*, 11(4):717–744.

Norman, D. A. (1983). Some observations on mental models. Mental Models. *Hillsdale, New Jersey, Erlbaum*, 7:14.

Ong, D. C., Zaki, J., and Goodman, N. D. (2019). Computational Models of Emotion Inference in Theory of Mind: A Review and Roadmap. *Topics in Cognitive Science*, 11(2):338–357.

Onishi, K. H. and Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science*, 308(5719):255–258.

Penrod, J. (2007). Living with uncertainty: concept advancement. *Journal of advanced nursing*, 57(6):658–667. Publisher: Wiley Online Library.

Peters, P. (2013). *The Cambridge dictionary of English grammar*. Cambridge University Press (CUP).

Pettigrew, T. F. and Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, 90(5):751. Publisher: American Psychological Association.

Piaget, J. (1976). Piaget's theory. In *Piaget and his school*, pages 11–23. Springer.

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

Preston, S. D. and de Waal, F. B. (2001). *Empathy: Its ultimate and proximate bases*.

Pyers, J. E. and Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological science*, 20(7):805–812. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Pynadath, D. V. and Marsella, S. (2007). Minimal Mental Models. pages 1038–1044.

Pynadath, D. V. and Marsella, S. C. (2005). PsychSim: Modeling theory of mind with decision-theoretic agents. In *IJCAI*, volume 5, pages 1181–1186.

Pöppel, J. (2023). Models for Satisficing Mentalizing.

Pöppel, J. and Kopp, S. (2019). Satisficing Mentalizing: Bayesian Models of Theory of Mind Reasoning in Scenarios with Different Uncertainties. *arXiv:1909.10419 [cs]*. arXiv: 1909.10419.

Pöppel, J., Marsella, S., and Kopp, S. (2021). Less Egocentric Bias in Theory of Mind When Observing Agents in Unbalanced Decision Problems. *Proceedings of CogSci, preprint*.

Qu, S. Q. and Dumay, J. (2011). The qualitative research interview. *Qualitative research in accounting & management*, 8(3):238–264. Publisher: Emerald Group Publishing Limited.

Ravenscroft, I. (2019). Folk Psychology as a Theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.

Reeves, B. and Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10).

Rizzolatti, G. and Fabbri-Destro, M. (2010). Mirror neurons: from discovery to autism. *Experimental Brain Research*, 200(3-4):223–237.

Rouse, W. B. and Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, 100(3):349. Publisher: American Psychological Association.

Rusch, T., Steixner-Kumar, S., Doshi, P., Spezio, M., and Gläscher, J. (2020). Theory of mind and decision science: towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488. Publisher: Elsevier.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., and Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5):1255–1266.

Sande, G. N. and Zanna, M. P. (1987). Cognitive dissonance theory: Collective actions and individual reactions. In *Theories of group behavior*, pages 49–69. Springer.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., and Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 19(2):65–72.

Schneider, D., Slaughter, V. P., and Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition*, 162:27–31. Publisher: Elsevier.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, 63(2):129. Publisher: American Psychological Association.

Simons, K. W. (1992). Rethinking mental states. *BUL Rev.*, 72:463. Publisher: HeinOnline.

Southgate, V. (2020). Are Infants Altercentric? page 76.

Spiller, S. A. (2011). Opportunity cost consideration. *Journal of Consumer Research*, 38(4):595–610. Publisher: University of Chicago Press Chicago, IL.

Tanovic, E., Gee, D. G., and Joormann, J. (2018). Intolerance of uncertainty: Neural and psychophysiological correlates of the perception of uncertainty as threatening. *Clinical psychology review*, 60:87–99. Publisher: Elsevier.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. page 10.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285. Publisher: American Association for the Advancement of Science.

Thagard, P. (2002). *Coherence in thought and action*. MIT press.

Thornton, M., Rmus, M., and Tamir, D. (2020). Mental state dynamics explain the origin of mental state concepts. *PsyArXiv.*

Todd, A. R., Forstmann, M., Burgmer, P., Brooks, A. W., and Galinsky, A. D. (2015). Anxious and egocentric: how specific emotions influence perspective taking. *Journal of Experimental Psychology: General*, 144(2):374–391.

Tononi, G. and Cirelli, C. (2014). Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81(1):12–34. Publisher: Elsevier.

Wandell, B. A. (1995). *Foundations of vision.* Sinauer Associates.

Watson, G. J. (2015). *"You shouldn't have to suffer for being who you are": An Examination of the Human Library Strategy for Challenging Prejudice and Increasing Respect for Difference*. PhD Thesis, Curtin University.

Westbrook, A. and Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, 15:395–415. Publisher: Springer.

Westra, E. (2019). Stereotypes, theory of mind, and the action–prediction hierarchy. *Synthese*, 196(7):2821–2846. Publisher: Springer.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraircing function of wrong beliefs in young children's understanding of deception. page 26.

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

Yongsatianchot, N. and Marsella, S. (2016). Integrating model-based prediction and facial expressions in the perception of emotion. *International Conference on Artificial General Intelligence*, pages 234–243.

Zainal, N. H. and Newman, M. G. (2018). Worry amplifies theory-of-mind reasoning for negatively valenced social stimuli in generalized anxiety disorder. *Journal of affective disorders*, 227:824–833. Publisher: Elsevier.

Zeng, Y., Zhao, Y., Zhang, T., Zhao, D., Zhao, F., and Lu, E. (2020). A Brain-Inspired Model of Theory of Mind. *Frontiers in Neurorobotics*, 14(60):1–17.

Zweig, A. and Weinshall, D. (2007). Exploiting object hierarchy: Combining models from different category levels. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE.