The mentalizing triangle: how interactions among self, other and object prompt mentalizing

Tian Ye

Thesis submitted to UCL for the degree of Doctor of Philosophy,

May 2020

I, Tian Ye, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed

Date

Abstract

To smoothly interact with other people requires individuals to generate appropriate responses based on other's mental states. The ability we rely on is termed mentalizing. As humans it seems that we are endowed with the abilities to rapidly process other's mental states, either by taking their perspectives or using mindreading skills. These abilities allow us to go beyond our direct experience of reality and to see or infer some of the contents of another's mental world.

Due to the complexity of social contexts, our mentalizing system needs to address a variety of challenges which put different requirements on either time or flexibility. During years of research, investigators have come up with various theories to explain how we cope with these challenges. Among them, the two-system account raised up by Apperly and colleagues (2010) has been favoured by many studies. Concisely, the two-system account claims that we have a fast-initiated mentalizing system which guarantees us to make quick judgments with limited cognitive resource; and a flexible system which allows deliberate thinking and enables mentalizing to generalize to multiple targets. Such a framework provides good explanations to debates such as whether preverbal young children can process mentalizing or not. But it is still largely unknown how healthy adults engage in mentalizing in everyday life. Specifically, why it seems easier for some targets to activate our mentalizing system, but with some others, we frequently fail to consider their perspectives or beliefs?

To give an explanation to this question, I adopted a different research orientation in my PhD from the two-system account, which considers the dynamic interactions among three key elements in mentalizing: the self, agent(s), and object(s). I put forward a mentalizing triangle model and assume the interactions in these triadic relationships act as gateways triggering mentalizing. Thus, with some agents, we feel more intimate with them, which makes it easier for us to think about their minds. Similarly, in certain context, the agent may have frequent interactions with the object, thus we become more motivated to engage in mentalizing. In the following chapters, I first reviewed current literatures and illustrate evidence that could support or oppose the triangle model, then examined these triangle hypotheses both from behavioural and neuroimaging levels.

In Study 1, I first measured mentalizing in the baseline condition where no interaction in the triangle relationships was provided. By adapting the false belief paradigm used by

Kovacs, Teglas, & Endress (2010), I imported the Signal Detection theory to obtain more indices which could reflect participants mentalizing processes. Results of this study showed that people have a weak tendency to ascribe other's beliefs when there is no interaction. Then, in Study 2, we added another condition which included the 'agent-object' interaction factor while using a similar paradigm in Study 1. Results in the noninteractiond condition replicated our findings of Study 1, but adding 'agent-object' interactions didn't boost mentalizing. Study 3 and 4 tested the 'self-agent' interaction hypothesis in visual perspective taking (VPT), another basic mentalizing ability. In Study 3, I adopted virtual reality approach and for the first time investigated how people select which perspective to take when exposed to multiple conflicting perspectives. Importantly, I examined whether the propensity to engage in VPT is correlated with how we perceive other people as humans, i.e. the humanization process. Congruent with our hypotheses, participant exhibited stronger propensity to take a more humanised agent's perspective. Then in Study 4, I used functional near-infrared spectroscopy (fNIRS) and investigated the neural mechanism underlying this finding. In general, the 'self-agent' hypothesis in the mentalizing triangle model was supported but not for the 'agent-object' hypothesis, which we consider may due to several approach limitations.

The findings in this thesis are derived from applying novel approaches to classic experimental paradigms, and have shown the potentials of using new techniques, such as VR and fNIRS, in investigating the philosophical question of mentalizing. It also enlights social cognitive studies by considering classic psychological methods such as the Signal Detection Theory in future research.

Impact Statement

This thesis investigated an important question in social cognition, mentalizing. The research explored the mechanism underlying mentalizing, especially when people engage in mentalizing processes such as theory-of-mind and visual perspective taking. Throughout four studies, the author proposed a mentalizing triangle model, and hypothesized that the interactions among the self, other and object in a mentalizing process constitute gateways towards spontaneous mentalizing behaviours. In a series of experiments, the author has used a number of novel methods and new techniques in psychological study, including the signal detection theory, Bayesian hierarchical model, virtual reality and neuroscience method -- functional near infrared spectroscopy (fNIRS). Such new approaches provided more information than traditional psychological measurements used in mentalizing research and aids to solve several critical debates ongoing in this field. In addition to the novel approaches used in the research, the author also followed practice advocated by the spirit of OpenScience and pre-registered most of the studies, to ensure the results are solid and objective. By virtual of these improvements, it is revealed that certain types of social interaction, especially in the self-other relationship, can prompt spontaneous mentalizing behaviours. The findings of the thesis are interesting and insightful, as it provides new perspectives to investigate how mentalizng behaviours differ in various contexts and with various targets. It also shed lights on research focusing on mentalizing abilities of individuals with autism, as their difficulties in attributing minds to neurotypical others may result from the difficulty in detecting the self, other and object interactions. It can also provide new thoughts to human-robot interaction research, benefiting the development of more socially interactive agents.

Acknowledgement

It's beautiful to be a human. We would like to think we are different from animals, that we are empowered with the ability to get to know other people's thoughts, feelings, intentions and beliefs, i.e. mentalizing. Through intangible media we communicate, often in nonverbal way, reading other's mind. So I would like to appreciate this mysterious cognitive ability in the beginning, which constitutes the main body of this thesis.

Back to life. I've never thought that I would finish this thesis in such a moment, when people all over the world are forced to stay away from their routine lifestyle because of the Covid-19, especially from the everyday social interactions, which is a bit ironic if you consider the topic of this thesis is about the core abilities in social cognition. But we know all of these are momentary. All these adversities will only make people want to get to know each other better, i.e. a stronger motivation to understand each other's mind☺. I feel a strong eager to meet my dear friends physically, including my respectable supervisor, Prof. Antonia Hamilton. The first reason to thank you is that you gave me such a good opportunity to become a member of the Social Neuroscience Group, to meet those lovely people. But most importantly, you walk me through the research field of social cognition, and are so supportive for my research career. You are a role model to me and I feel honoured to have you as my PhD supervisor.

I would also like to thank my second supervisor, Prof. Steve Fleming, who guided me to the world of Bayesian analysis, which I believe will bring me tremendous benefit in the future.

Special thanks to my dear friend, best PhD buddy, Dr. Roser Cañigueral. I have millions of reasons to say thanks. Without you, it would be much less funny and much harder both for my everyday life and study. Our every trip together has become the best memories in my life and I'm sure we will create more happiness together in the future.

I also want to thank Dr. Alexandra Georgescu and Dr. Paola Pinti, because you are both good friends and teachers to me. Thank you for always being generous to share you experience with me and putting up with my endless questions. I'm also so thankful for having so much fun with you no matter when we tripped together or in leisure time. I feel so lucky to know you two, such amazing persons!

I also would like to say ‘thank you’ to my other labmates, Sujatha Krishnan-Barman, Patrick Folk, Alexis Macintyre, Sara de Felice, Isla Jones, Anna Ciaunica and Paul Forbes, and many other colleagues for giving me so much help.

Of course, I need to thank my dear Chinese friends in London. Cai Qing, Liu Junfei, Wu Ruihan, Cai Peijun, my roommate Lu Yi, Yin Feng, Dr. Ma, Li Heyu, Hu Haochen, Li Yanzhe. I know many of you may never read this thesis. But you all give my PhD life so much fun and in fact, some ideas of this thesis came from our interactions☺.

Lastly, but most importantly, I want to thank my parents, especially my mom. You are my spirit for pursuing more in life, including answers to unknown questions in research. Thanks for your unconditional love, I will also love you forever.

Tian Ye

26th, May, 2020

Contents

Figures

Tables

Chapter 1. Introduction

Consider this situation. Elaine and Kelly have been good friends for years. One day they were shopping in a clothes store, Kelly saw Elaine looking at the shop mannequin, which was wearing a beautiful fine-cut dress. The dress looked like a good fit for Elaine and it seemed that she liked it a lot. Then they saw the price tag and Elaine hesitated and decided to look around. Kelly knew that Elaine loved this dress but was just worried about the price. So that while Elaine was in the fitting room, Kelly secretly bought the dress and put it in her bag with other clothes, planning to give it to Elaine days later on her birthday.

Living a social life, we are quite familiar with scenarios as above. To become a kind friend, a capable group leader or a considerate social worker, sometimes it is necessary for us to go beyond other's explicit words or behaviours, but rather make inferences about what they see, feel, believe or desire, so as to predict their behaviours. For example, in the above episode, although Elaine said she wanted to look around, Kelly knew at once that indeed she liked the dress but was only worried about the price, since she changed her mind after noticing the price tag. Thus, by linking what others see and what they think, we sometimes can perceive those intentions hiding behind a person's explicit words or behaviours, such that we are able to react based on this truly useful information.

The psychological term for understanding other's mental content is mentalizing, which is a building block for normal social life. Although understanding others' minds is still challenging for modern artificial intelligence, as humans, it seems we are endowed with such mentalizing systems which enable us to rapidly detect other's thoughts and feelings. People can quickly infer other's feelings from their eye-gaze (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), reason their intentions from their biological movements (Pelphrey, Morris, & McCarthy, 2004; Sartori, Becchio, & Castiello, 2011), and more comprehensively, interpret beliefs from their interactions with the surroundings (Senju, Southgate, White, & Frith, 2009). It is worth noting that in most cases these processes take place in a nonverbal manner.

The extent to which people are good at mentalizing has been found to closely relates with their social competent behaviours and interpersonal relationship (Repocholi & Slaughter, 2004), to the extreme even with behaviours of manipulative purposes (e.g. Machiavellian behaviours) (Lyons, Caldwell, & Shultz, 2010; Whiten, 1997). In contrast, individuals with autism usually find it difficult to interact with neurotypical people, and such difficulties are related to their distinct behavioural patterns on some mentalizing tasks, which target especially on theory-of-mind (ToM) and visual perspective taking (VPT), two core abilities in mentalizing (Baron-cohen, Leslie, & Frith, 1985; Baron-Cohen, 2000).

In the following paragraphs, I shall first introduce the concepts and current results in ToM and VPT research, then give a brief introduction to the two-system account and its limitations. In the third part, I suggested a triangle model for mentalizing, which focuses on the dynamic interactions among self, agent (s), and the object(s). The last part of this chapter will focus on current methodological challenges and provide suggestions for improvements.

1.1 Theory-of-mind (ToM) and visual perspective taking (VPT)

When thinking about what other people are thinking, we usually need to temporarily step out from our own psychological world, and infer other's thoughts by 'standing in their shoes'. In this process, visual perspective taking (VPT) refers to switching our perspective and theory-of-mind (ToM) is the process we used to reason about other people's beliefs, thoughts or feelings. For example, in the opening shopping example, Kelly bought the dress while Elaine was in the fitting room. This is because Kelly knew that if Elaine didn't see her buying the dress (the VPT process), then she wouldn't know what her present is (the ToM process). Although in daily life the 'seeing-knowing' rule naturally linked VPT and ToM close to each other, in laboratory research, researchers are used to investigate them separately by using distinct paradigms. So in the following paragraphs, I will first introduce the concept and current research of VPT and ToM separately, then provide a brief summary of the neuroimaging findings on their neural mechanisms.

1.1.1 Theory-of-mind (ToM)

The concept of ToM dates back to 1978, when Premack and Woodruff (1978) first put forward the question '*Does the chimpanzee have a theory-of-mind?*' Their original purpose was to figure out whether the chimpanzees' ability to solve a series of interaction problems involves considering their conspecifics' mental states. This study later powered the research on humans' theory-of-mind ability, and researchers started asking when humans acquire such ability, and how it relates to our social behaviours in real life. It was soon revealed that not everyone is equipped with fully-developed ToM ability. For example, children before 4-year-old, although they are capable to express themselves fluently, they tend to predict other's behaviours based on the reality rather than what other people currently believe (Wimmer & Perner, 1983). Likewise, impairment on social interaction skills has been recognized as a core deficit for individuals with autism (DSM -V), with studies showing that they tend to make predictions of other's behaviours based on reality, while neglecting their current beliefs (Baron-Cohen, Leslie, & Frith, 1985).

Such reality-based predictions revealed a key understanding possibly impaired among young children and autistic individuals, which is termed 'false belief'. As typical adults, we understand that in most cases people's reactions to reality is based on the beliefs they hold. For example, Elaine might return to the store to get the dress in early September because she believed it would be on sale in autumn. But beliefs are not always reflections of reality. False belief refers to those beliefs that contradict reality, so that if an individual can explain or predict other's behaviours based on their false beliefs, it suggests that he is capable of representing the other's mental states separately from reality. Thus, researchers agree that the understanding of 'false belief' is a benchmark in the development of ToM ability.

1.1.1.1 The false belief task

In laboratory research, the false belief task (FBT) is a widely-used task to test the emergence and individual differences of ToM. The classic FBT involves two characters: a little girl named Sally and another named Anne (Figure 1-1.A). Sally first put her favourite toy in a basket then left. While her absence, Anne moved the toy to a box, and participants were usually asked where Sally would look for the toy upon her return (Baron-Cohen et al., 1985). The rationale behind this task is that individuals with fully-developed mentalizing ability will be able to predict Sally's behaviour based on her current belief (that she will look for the toy in the basket), rather than the reality at that specific moment (the toy is in the box). Using a narrative version of this task, researchers found that children from 4-year-old start to take into account of other's mental contents when predicting their behaviours (Wimmer & Perner, 1983), and such a characteristic event is usually regarded as 'a radical shift in their understanding of the mind', since now children seem to have acquired a 'representation' ability, that they become aware of both their own beliefs states and are also able to attribute such states to others (Bloom & German, 2000).

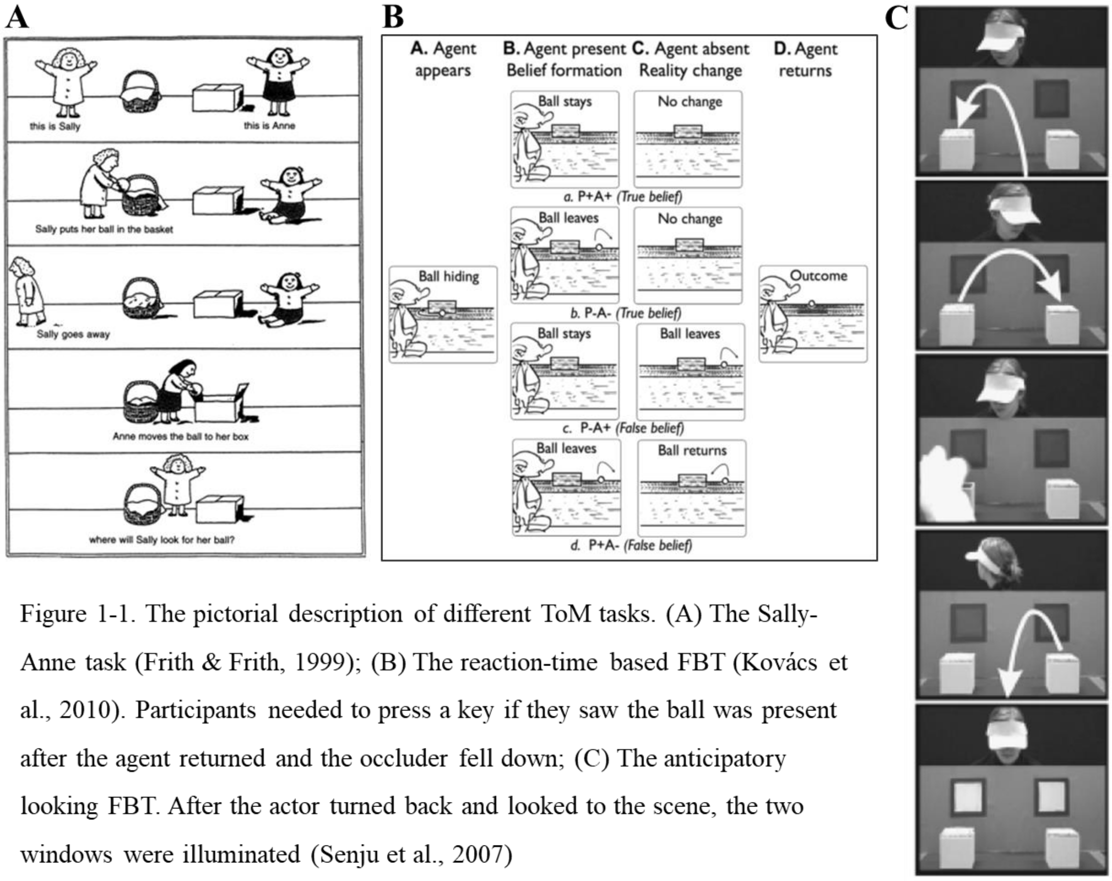


Figure 1-1. The pictorial description of different ToM tasks. (A) The Sally-Anne task (Frith & Frith, 1999); (B) The reaction-time based FBT (Kovács et al., 2010). Participants needed to press a key if they saw the ball was present after the agent returned and the occluder fell down; (C) The anticipatory looking FBT. After the actor turned back and looked to the scene, the two windows were illuminated (Senju et al., 2007)

1.1.1.2 Behavioural signs of belief reasoning

Despite the significance of Sally-Anne task in studying ToM, some researchers argued that the late development of ToM may relate to the testing methods, specifically, young children's performance on this task may be constrained by their language comprehension abilities and other inadequate general cognitive abilities (e.g. executive function and working memory)(Onishi & Baillargeon, 2005; Kampis, Parise, Csibra, & Kovács, 2015; Perner & Roessler, 2012; Baillargeon, 2010). In other words, children younger than four or even infants may be already equipped with an understanding of other's mental states, yet they are unable to use this ability due to insufficient support from other cognitive constructs. Following this hypothesis, researchers started to design non-verbal versions of FBT and seek to gather evidence for the early emergence of ToM.

A classic research approach to study infants' cognitive ability is the violation-of-expectation (VOE) paradigm. Infants and toddlers are too immature to use language; but when seeing surprising events, or consequences that violate their expectations, they tend to show prolonged looking time. Onishi and Baillargeon (2005) tested with this method whether

15-month-old infants possess the ability to predict other’s behaviour based on beliefs. Infants watched an actor interacting with a target toy, which can be hidden in a green or yellow box. In a familiarization trial, they first learned that the actor would always reach the toy in the end. Then they underwent a single test-trial where the position of the toy changed, and critically depending on the experimental conditions, the actor saw (true belief) or didn't see (false belief) the change. They found that in both true belief and false belief conditions, infants looked longer to trials where the actor reached the toy location which violated her belief. So the researchers argued that infants can realize that others’ acts are based on their beliefs, although sometimes the beliefs are against reality.

Southgate, Senju and Csibra (2007) tested another form of looking behaviour with infants by introducing an eye-tracking approach to Onish and Baillargeon’s task (see Figure 1-1.C). Participants were familiarized with the events that an agent would reach an object at the end of each video. During the testing trial, the object was moved to a different location while the agent was looking away, and researchers then recorded participant’s looking behaviour after the agent turned back looking to the boxes (Senju et al., 2009; Southgate et al., 2007). They found that both neurotypical adults and children turned to look towards the position where the agent believed the toy is. Such looking behaviour takes place before the agent reached the toy, thus suggested that participants were anticipating what was going to happen. Therefore, they name their task anticipatory looking paradigm.

Using Senju’s paradigm, Schneider recruited adult participants and further revealed that such anticipatory looking behaviour can still be observed when participants did this task for a prolonged time period (Schneider, Slaughter, Bayliss, & Dux, 2013). Moreover, Schneider manipulated the instructions given to the participants and found that anticipatory looking persists when the task required participants to perform a belief-irrelevant task (Schneider, Nott, & Dux, 2014), but it disappeared when participants were asked to perform a dual-task. These results may suggest that anticipatory looking in the FBT task at least minimally relies on general cognitive ability. But there is another possibility that it may be not reliable to use looking pattern as the sole measurement of mentalizing, since people may still be able to represent other’s beliefs under cognitive load, however simply do not show it in their gaze patterns.

More recently, investigators started using reaction time (RT) as the main index to test the influence of belief reasoning. Kovács designed an object-detection FBT (see Figure 1-1.B) for testing both adult and infant participants. In their paradigm, a Smurf placed a ball on a table, which then moved behind an occluder. After a short period, the ball moved out from the occluder then either moved out of the scene, or moved back behind the occluder. Due to different manipulations, the Smurf either saw the ball's full movements, or didn't see because it left the scene. At the end of each video, the Smurf looked towards the occluder, and adult participants were instructed to press a key as fast as possible when they saw a ball after the occluder disappeared (Kovacs et al., 2010). Kovács found that in conditions where the Smurf believed the ball was behind the occluder, no matter this belief was true or false, participants responded faster to the presence of the ball. With the infant participants, they analysed infants gaze behaviour by using a similar approach with the VOE method. Infants' gaze pattern also revealed longer gaze time when the consequence violates the Smurf's belief, even when the belief was false.

But Kovacs' findings with adult participants were later challenged by Phillipes et al. (2015). They conducted 13 experiments using Kovacs' RT-based false belief task, and concluded that the RT difference found between different beliefs conditions was due to the attention check manipulation in Kovacs' task. Since participants needed to press a key indicating they attended to the videos at different time points in the four belief conditions, their reaction time for detecting the target ball in the end of each video was influenced by the refractory movements. The inconsistency between Kovacs' and Phillipe's results again suggest that using only one index as the measurement of mentalizing can be easily affected by belief-irrelevant factors.

1.1.1.3 A short summary of current research on ToM

Results of current research on ToM suggest that ascribing mental states to other people is not a robust process. It may subject to the influence coming from both internal and external factors, such as cognitive load, belief-irrelevant events and social contexts. However, since current studies usually rely on a single behavioural index (e.g. reaction time or gaze-pattern) for analyses, the results are thus easy to be changed which causes debates. We assume that spontaneous mentalizing may exist given proper social context, however such process may manifest in distinct cognitive processes such that it may change different behavioural indices. For example, representing other's beliefs may drive people to attend more to the target which was noticed by others, and thus changing anticipatory looking, but it could also influence our memory thus we can remember them better. In this way, it points out the necessity of using different behavioural indices to comprehensively measure mentalizing behaviours, so as to better reveal how it changes in multiple social scenarios.

1.1.2 Visual perspective taking

Compared with ToM which usually deals with abstract concepts (e.g. beliefs), VPT focuses on more concrete contents. It refers to the ability we have to infer other's visual experience, usually by switching our viewpoint to imagine how a scene appears to another person. Previous research revealed there are two levels of VPT: level-1 VPT calculates if another person can see the same target as we do or not and level-2 VPT infers how the target looks different from another viewpoint (Flavell, 1977). These two levels of VPT not only differ on the contents they specify, also on their developmental trajectories. The knowledge of whether others can see something or not emerges as early as 14-month-old (Sodian, Thoermer, & Metz, 2007). On the contrary, studies showed that children before 4-year-old usually have difficulty in working out other's perspective if the same object appears differently to another person (Moll & Meltzoff, 2011; Moll, Meltzoff, Merzsch, & Tomasello, 2013). In the following paragraphs, I will first compare the paradigms of VPT widely-used in social cognitive research, and then summarize current findings and their indications.

1.1.2.1 Research approaches of VPT

The psychological perspective taking research starts from the old *Three Mountain Problem* (Piaget & Inhelder, 1948), which in the beginning aimed to probe children's spatial perception. Most of the recent research now has been conceptualized within Flavell's (1977) distinction between level-1 and level-2 VPT. A typical example is the 'altercentric intrusion' effect, which refers to the phenomenon that a quasi-automatic processing of other's perspective sometimes interferes the processing of our own visual inputs (Samson, Apperly, Braithwaite, Andrews, & Scott, 2010; Nielsen, Slade, Levy, & Holmes, 2015). Samson investigated whether leve-1 VPT can occur in a spontaneous or even automatic way. In their study, participants were requested to report the numbers of discs on the walls with a virtual character facing one of the walls (see Figure 1-2.A). Critically, in some trials one or two discs appeared on the wall behind the character, thus participants can see a different number of discs than the character. By using this paradigm, Samson found that participants were less accurate and responded slower to the stimuli when there was an incongruent perspective, thus provided the first evidence for an automatic account of level-1 VPT.

The altercentric intrusion effect was then further evidenced in a number of other studies.

Surtees & Apperly (2012) tested Samson’s (2010) paradigm with both children (6-10 years old) and adults, they replicated Samson’s results by showing both children and adults were slower in reporting the number of discs when there was an incongruent perspective, however such interference effect disappeared when the virtual agent was replaced by a non-social stick. MacDorman and colleagues (2013) revealed that the altercentric intrusion effect persists with a variety of virtual agents. Other studies suggested that automatic level-1 VPT exists among different populations, including people suffering from alcohol-dependence and individuals with autism (Cox, Chandler, Simpson, & Riggs, 2016; Schwarzkopf, Schilbach, Vogeley, & Timmermans, 2012). Together results from these studies indicate that level-1 VPT seems to be immune to a majority of social and non-social factors. Based on these findings, some researchers claim that level-1 VPT is an involuntary, or even obligatory process, i.e. it can be initiated whenever we encounter a social agent and is not modulated (Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010; Furlanetto, Becchio, Samson, & Apperly, 2015; Todd, Cameron, & Simpson, 2017; Troje, 2013; Capozzi, Cavallo, Furlanetto, & Becchio, 2014).

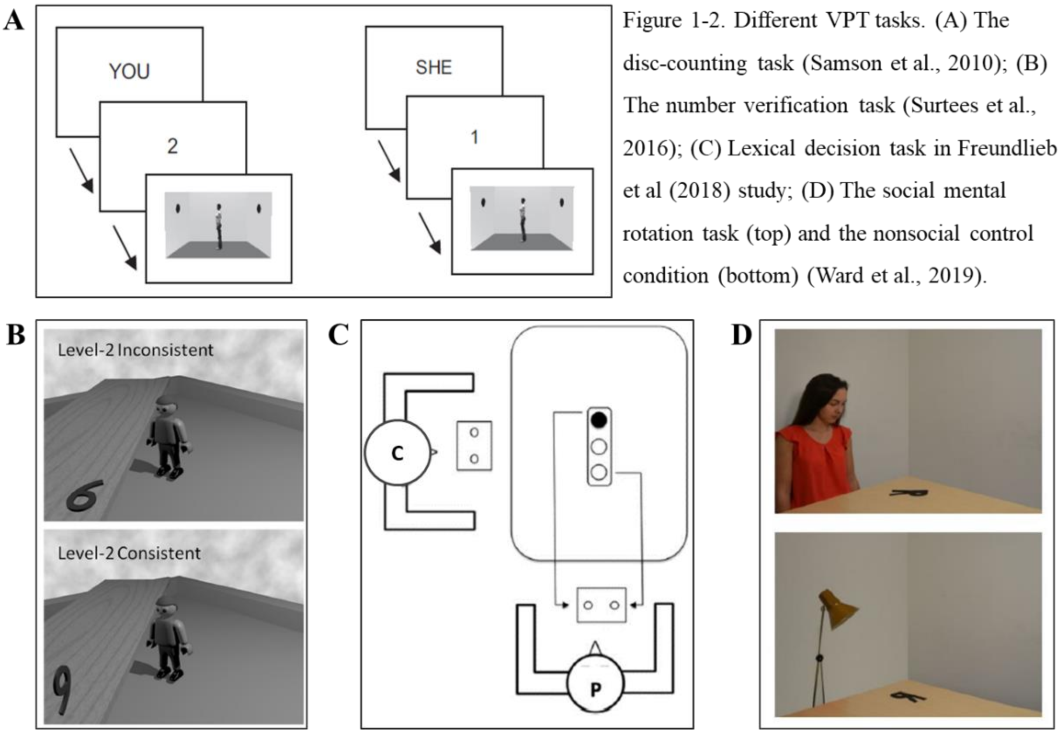


Figure 1-2. Different VPT tasks. (A) The disc-counting task (Samson et al., 2010); (B) The number verification task (Surtees et al., 2016); (C) Lexical decision task in Freundlieb et al (2018) study; (D) The social mental rotation task (top) and the nonsocial control condition (bottom) (Ward et al., 2019).

Despite a large body of research supporting the automatic view of level-1 VPT, several others questioned if such ‘altercentric intrusion’ indeed reflects a social cognitive effect because a similar results pattern has also been observed for non-social targets. Santiesteban, Catmur, Hopkins, Bird & Heyes (2014) compared the interference effect from a virtual agent and that from an arrow, and they revealed similar effects between the two conditions. Results from this study suggest VPT is likely to be controlled by a domain general process, which does not distinguish between social or non-social stimuli. However, a recent study from Nielsen and colleagues (2015) systematically manipulated the social essence of different agents in Samson’s task, and discovered that although the altercentric intrusion effect existed in all conditions, with a human agent a stronger effect can be observed. Crucially, Furnelatto et al. (2015) tested this effect with the same social agent wearing or not wearing a blindfold, and only in the condition where the virtual agent could see was the participant’s performance impaired. These recent results thus indicate a genuine automatic level-1 VPT account.

Corroborating results from Samson’s paradigm, Freundlieb and colleagues (Freundlieb, Kovács, & Sebanz, 2016; Freundlieb, Sebanz, & Kovács, 2017b) revealed spontaneous level-1 VPT when participants were performing a simple stimulus-response (SR) compatibility task (see Figure 1-2.C). Participants and a confederate sat 90 degree nearby a table, and needed to respond to a top black disc using either the left-index or right-index finger. They found that when participants’ spatial response layout mapped with the confederate’s orientation, their reaction time was facilitated. But such results were only revealed in conditions when the confederate was perceived as an intentionally acting agent, not when the confederate’s vision is blindfolded. Again, such results supported the theory that automatic level-1 VPT is a social process rather than being controlled by domain general processes.

Contrary to level-1 VPT, studies on level-2 VPT provided many conflicting results. Researchers found when making lexical decisions, spontaneous VPT occurred and facilitated the recognition of letters and words. Freundlieb and colleagues found that participants were much quicker to recognize a word when it orients to a person in a semantic categorization task (Freundlieb, Kovács, & Sebanz, 2018). Similarly, Ward et al. also discovered an RT boost effect by using a social mental rotation task (Ward, Ganis, & Bach, 2019; Ward, Ganis, Mcdonough, & Bach, see Figure 1-2.D). In their study, participants completed a classic mental rotation task where a letter ‘R’ or ‘Я’ is presented in different orientations and the participant must judge if it is a normal or mirror reversed letter. Critically, a confederate was sometimes visible on the left or right of the table. When the confederate was not present, they replicated the classic mental rotation effect, that items required more mental rotation were recognized slower. Critically, when there was a confederate and when the letters were rotated away from the participant but were oriented towards the confederate, they were also recognised more rapidly by the participants, suggested that participants ‘stand-in’ the confederate’s position when completing the task.

Unlike results from the lexical paradigms, results with a number-verification paradigm suggest that involuntary level-2 VPT may be conditional and rely on other cognitive processes. In one study, Surtees asked participants to recognize several numbers with a cartoon figure standing opposite to them (Surtees, Samson, & Apperly, 2016, Figure 1-2.B). In each trial a cue first showed up on the screen (‘six’ or ‘nine’), then a number 6 or 9 came up either in an upright (congruent perspective, participants and the agent saw the same number) or flat manner (incongruent perspective, e.g. participants saw ‘6’ but the agent saw ‘9’) in between participants and the cartoon figure. Participants were instructed to report whether the cue correctly matched up with the current stimulus based on their own or the agent’s perspective. Results revealed the altercentric intrusion effect when there is an incongruent perspective. But unlike in level-1 VPT tasks, such interference only was only observed in mixed blocks where participants needed to switch between self and other perspectives from trial to trial, thus suggesting that specific experimental contexts might have changed the way participants processed other’s perspectives.

Several follow-up studies further suggested that involuntary level-2 VPT is situational rather than universal. Such situational factors include but maybe not limited to cognitive load, social motivations or background knowledge. For example, Elekes adopted the number verification task and found spontaneous level-2 VPT only occurred with the partner was processing the object’s perspective dependent feature (Elekes, Varga, & Király, 2016a). Such a result is similar to findings reported in the lexical task, in which spontaneous level-2 VPT was found only when participant knew that the partner was attending to the same target (Freundlieb et al., 2018). Cane et al. also reported that spontaneous level-2 VPT can be modulated by motivational factors such as monetary reward (Cane, Ferguson, & Apperly, 2017). These findings suggest that level-2 VPT can occur spontaneously or involuntarily, but participant’s involvement in other’s perspective can also be adjusted. Thus, unlike level-1 VPT which usually takes place rapidly, calculating how other people see something (level-2 VPT) may rely on several social or non-social prerequisites.

1.1.2.2 Implications of current findings on VPT

Consistent with results from ToM studies, research on VPT also suggests there are different subcomponents of this ability. Akin to non-verbal or implicit theory-of-mind, there is a rapid perspective taking process, which enables people to know what others can see. Similarly, a slow, more flexible level-2 VPT calculates how something appears to other people, which is like the explicit theory-of-mind process that provides us detailed information about other's mental state. Complementary to each other, these two processes address different issues in real life, endow us the ability to effectively and efficiently interact with others.

However, beyond this two-system account, there is another possibility that both theory-of-mind and the VPT process are regulated by other factors, which can change the propensity to engage in processing other's mental state. Such factors act like gates controlling people's mentalizing abilities, and might determine whether or not people engage in mentalizing about a particular target person. In the second and third part of this chapter, I will elaborate these ideas and point out several potential factors which control the gate towards mentalizing.

1.1.3 The neural mechanisms of mentalizing

To further understand the mentalizing process, researchers started to ask what brain activities take part in the process of thinking about other's mental states. Recent findings suggest a large overlap between brain areas involved in various mentalizing tasks, which includes the engagement from the temporal-parietal junction (TPJ), medial prefrontal cortex (mPFC), the superior temporal sulcus (STS) and the precuneus (Frith & Frith, 2006; Saxe & Kanwisher, 2003; Samson, Apperly, Kathirgamanathan, & Humphreys, 2005; Bukowski, 2018; Dumontheil, Küster, Apperly, & Blakemore, 2010; Schurz et al., 2015). In particular, many studies have discovered that the bilateral temporal-parietal junction selectively involved in processing other's beliefs, and tasks engaging thinking from other's perspective, with regions of mPFC and STS playing a more supportive role (Hyde, Betancourt & Samson, 2015; Saxe & Kanwisher, 2003).

Fletcher et al. first used positron emission tomography and investigated brain activity when participants were performing a story-based false belief task. They found that processing other's mental states significantly increased cerebral blood flows in bilateral temporal poles, left superior temporal gyrus and the prefrontal cortex (Fletcher et al., 1995). Saxe and Kanwisher (2003) investigated brain activities using functional magnetic resonance imaging (fMRI) when participants were reading short stories involving or not involving another person's mental states. They found that a region within the TPJ bilaterally showed greater activity when participants need to think about other's goals or beliefs, compared with when stories were about nonhuman objects. Further, by using more stringent control conditions they further clarified the selective role of TPJ in mentalizing (Saxe & Powell, 2006). Participants read stories about other's thoughts, bodily sensations or appearance in the scanner. Results showed that although bilateral TPJ, posterior cingulate and medial prefrontal cortex were all involved in these activities, bilateral TPJ and posterior cingulate selectively responded to reading stories about other's thoughts. Similar results were also found when nonverbal false belief tasks (FBT) were used. For example, Sommer and colleagues (2007) adopted a pictorial FBT and found that when asking participants to explicitly judge whether the story outcomes were expected or not based on the protagonist's beliefs, stronger activation was found in the rTPJ in the false belief condition.

But it is still debated whether TPJ is involved in spontaneous or implicit false belief ascription. Kovács et al. (2014) tested the RT-based FBT task (Kovács et al., 2010) in the fMRI scanner, results revealed a significant increase in right TPJ activation when the character was led to falsely believe the ball was present, compared with conditions when its belief was consistent with reality. However, another fMRI study using the anticipatory looking FBT revealed different results. In this study, people were asked to watch the video clips freely (spontaneous mentalizing conditions), but results failed to show that rTPJ had stronger activities in the false belief conditions compared with the true belief conditions (Schneider, Slaughter, Becker, & Dux, 2014). Using the same paradigm as Schneider et al. (2014) but with a functional near-infrared spectrum (fNIRS) method, Hyde et al. (2015) found a significant difference in TPJ responses between false belief and true belief conditions. They assumed their results difference compared with the previous study (Schneider et al, 2014) may due to the higher temporal resolution of fNIRS (50 Hz) compared with fMRI (0.5-1Hz), which allowed researchers to capture the moment-to-moment conceptual change in the stimuli.

Research on VPT also reported large overlap of neural substrates with that of ToM, with a highlight on the role of TPJ. Ramsey and colleagues used fMRI and found that both TPJ and IPL related areas are involved in perspective taking in the disc-counting task, but the frontoparietal cortex showed stronger activity when inhibition of an egocentric perspective was needed (Ramsey, Hansen, Apperly, & Samson, 2013). Other studies mainly used electrophysiological approaches. For example, Beck and colleagues recorded participants' electroencephalography (EEG) signals using a frequency-tagging approach when participants were performing the disc-counting task (Beck, Rossion, & Samson, 2018). They presented participants pictures from the disc-counting paradigm at a frequency of 2.5Hz, but the pictures with incongruent perspectives were presented at a frequency of 0.5Hz. Thus, they are able to distinguish the neural signal coupling with perspective change. Results showed that a neural signature over the central and prefrontal electrode sites specifically focuses on the perspective-change, critically, they claimed the signal was possibly due to the processes generated in the TPJ areas. By using event-related potential (ERP), McCleery and colleagues found that a middle-latency temporalparietal component (TP450) is sensitive to perspectives. Particularly, it demonstrated the longest latency when participants were required to report from an inconsistent altercentric perspective. Although both studies indicated the significance of TPJ during VPT tasks, considering the poor spatial resolution of EEG/ERP, more evidence is still needed from well-designed neuroimaging studies.

1.2 The two-system mentalizing account

Based on previous behavioural and neuroimaging results, some researchers have begun to use a two-system theory to encompass different mentalizing processes. In general, the two-system theories claim that the general mentalizing process is composed of two subsystems, but different researchers came up with various descriptions of their two systems. Among them, Apperly provided the most detailed and systematic illustration(Apperly & Butterfill, 2009; Butterfill & Apperly, 2013)

Apperly hypothesized that mentalizing is governed by an implicit and explicit process. His arguments started from a fundamental conflict when mentalizing is used to solve interaction problems: how to balance the need from flexibility and efficiency. Returning to our opening example, when Kelly saw Elaine looking at the mannequin, almost automatically she knew that Elaine was looking at the dress; it is very unlikely that Kelly would care about what the mannequin was ‘looking’ at. However, when Elaine told her she didn’t want to buy the dress, it may require Kelly to think and relate to the previous fact that they saw the price of the dress, to figure out Elaine’s true intention.

To address this conflict, Apperly put forward an implicit mentalizing system, which is likely to operate in an unconscious, unintentional, attentionally efficient and perhaps automatic way (Apperly, 2009; Baillargeon et al., 2010; Clements, 1994; Onishi & Baillargeon, 2005; Low & Perner, 2012; Schneider et al., 2017). Compared with the explicit mentalizing system, the implicit system has superficial representations of other’s mental contents, however requires less cognitive resources and relies less on verbal skills. Since that it relies on limited cognitive resource to operate, it has the potential to address mentalizing tasks where prompt social responses are needed during interactions, for instance, to calculate if the agent can see a different number of discs in Samson's task (Samson et al., 2010).

In contrast, the explicit mentalizing system is more flexible and can be applied to different targets. It usually operates in an offline manner, and requires an individual to understand the propositional relationships in belief reasoning. As it usually unfolds in a more elaborative manner, theories also claim that the explicit system cooperates closely with several domain general processes such as working memory and executive function (Mckinnon & Moskovitch 2006; German & Hehman, 2006).

Apperly's two-system view is important in two ways. First, with a separation between the implicit and explicit systems, we can give better explanations on occasions where different behavioural indices support different hypotheses. van der Wel (2014) instructed participants to do an explicit FBT task, where participants must track either their own or the agent's belief, and moved the cursor to a target ball at the end of the video. They found that in the 'self' condition, although participants are correct on locating the target ball, their movement trajectories were affected by the agent's false belief, indicating the dissociative explicit and implicit mentalizing processes supervise different behaviours in the same task.

Second, Apperly argues that the implicit mentalizing system is shared by both children and adults. Thus, it is possible to explain why young children before 4-year-old, who are unable to provide correct verbal answers to belief questions in the FBT, however showed anticipatory looking as a sign of understanding beliefs. As the implicit system needs limited domain-general cognitive resource, it is even possible that some primates such as chimpanzees also share this system, such that they are able to fulfil some basic social cognitive process to cooperate (O'Connell & Dunbar, 2003; Martin & Santos, 2013; Krupenye, Kano, Hirata, Call, & Tomasello, 2016).

However, several studies questioned the existence of the implicit mentalizing system, as much of the supporting evidence is still hotly debated. For example, in Kovács' study, when the Smurf has a false belief that the ball was behind the occuluder, participants were faster on detecting the presence of the ball. But as mentioned above, using the same paradigm, Philips (2015) argued that the RT boost effect in Kovács' study might be due to a 'keyhit' event embedded in different timepoints under different conditions, which is irrelevant to belief reasoning. Heyes (2014) also asked whether the difference found in previous studies reveal a genuine social cognitive process, or 'submentalizing', which indicates that domain-general cognitive processes simulated such an effect. Taking theh altercentric intrusion effect as an example, reserachers questioned whether same data could be explained by a domain general attention shifting effect (Santiesteban, Catmur, Hopkins, Bird, & Heyes, 2014). Results from the anticipatory looking paradigm suffered from debates on the entangling relationship between belief-reasoning, selective attention and response-inhibitory in this task (de Bruin, Strijbos, & Slors, 2011). For example, Heyes claimed that neurotypical participants looked more to the location registered in the agent's false belief because they were distracted by the agent's head turning. In contrast, individuals with autism showed no difference in looking pattern because they were less distracted, instead of being impaired on the implicit mentalizing system (Heyes, 2014). Until now, all these debates are still unresolved.

I believe the questions in hot debate are critical to the research on mentalizing however it is extremely difficult to solve these questions ultimately. But beyond these theoretical debates, recently researchers have pointed out there are other limitations of the two-system account. The first is that it might be oversimplified when explaining mentalizing issues in real life (Christensen & Michael, 2016). Although typical adults are usually regarded as equipped with full-blown mentalizing abilities, it is common that we do not always engage in mentalizing whenever a social agent is encountered (which is a hypothesis based on the implicit mentalizing account). For example, when waiting for a train at the train station, it would be impossible for us to track the visual contents of all the dozens of people passing by; instead, we may only take the perspective of our friend. So here we hypothesize that the propensity people engage in mentalizing may be dependent on different targets and contexts, as they provide useful clues for us to judge the 'self-agent' and 'agent-object' relationship. Take the shopping case in the beginning as an example. Kelly may not be able to report what a stranger in the shop was looking at if she is suddenly asked, as the level-1 VPT process was not switched on for the stranger. But if the stranger suddenly reaches for the same dress which Kelly likes, it would be more likely for Kelly to work out what can be seen by the stranger. Such real-life scenarios indicate that mentalizing is very likely to be target and context-specific. The other limitation is that in Apperly's theory, the two systems barely interact with each other. However, neuroimaging studies showed there are large overlap between neural substrates in charge of these two processes, indicating great similarities between the neural mechanism of these two processes.

Addressing these limitations is important first because it can bridge the laboratory findings with real-life mentalizing issues. Moreover, with more mentalizing-related factors identified, research can proceed to profile how mentalizing is integrated in the overall cognitive system, thus is helpful for promoting research in other fields, such as designing socially intelligent virtual agents. Thus, in this PhD thesis, I am not focusing on solving the theoretical debates between the two-system account and other theories, but rather I am interested in exploring the enabling factors of mentalizing behaviours, which has been omitted by the majority of studies for a long time. To explore when and how spontaneous mentalizing behaviour could happen can help us provide explanations to why we infer other's mental states differently when encounter different agents, and with different targets.

1.3 The mentalizing triangle

In this thesis, I proposed a new model which addresses two questions: 1) what triggers the mentalizing process, in laboratory tasks or in real life; 2) what makes people invoke mentalizing differently with different agents? To address these problems, I start by taking a step back and consider what shapes social interaction. A typical mentalizing issue is usually compose of three elements: self, agent(s), and the object(s) (see Figure 1-3). That is, every mentalising study involves a participant (self) thinking about how an agent (other) views or represents or understands an object (target). Studies differ in how or what takes these different roles. The agent(s) can be a real person or an imaginary other. Similarly, the object(s) can be a concrete object or an abstract proposition. For instance, in experimental paradigms, the other agent can be a real person (Freundlieb et al., 2016; Schneider et al., 2014), a cartoon figure (Kovacs et al., 2010), or animated geometric figures (Abella, Happe & Frith, 2000). Accordingly, the target object can also be a ball (Kovacs et al., 2010) or a number (Surtees et al., 2016). These three core components constitute the mentalizing triangle (Figure 1-3).

Here, we propose that the relationships among these three items are important (sides of the triangle), and that close links between all three items promote mentalising. In contrast, if we encounter other people or objects without linking them together into a 'mentalising triangle', we might not engage in mentalising. Thus, the relationships embedded in the sides of the triangle are critical, and each might be part of a gate controlling the propensity to engage in mentalizing. Before different mentalizing processes are initiated, such relationships were first perceived and evaluated. Depending on an individual's internal states and the current task, distinct features of such relationships modulate the propensity of people to engage in mentalizing.

This model predicts that, in the shopping example, Kelly is more likely to take the perspective of her friend Elaine instead of the mannequin or a shop assistant because she has a closer ‘self-agent’ relationship with Elaine. Similarly, she is more likely to take perspective about who can see the dress that she or Elaine likes, compared to the trash can in the corner. Although these examples are common in life, they may provide critical information indicating that people selectively use their mentalizing ability according to the dynamics in the mentalizing triangle. In the following paragraphs, I will focus on explaining how such features in this triplet can potentially change the propensity of people to engage in mentalizing.

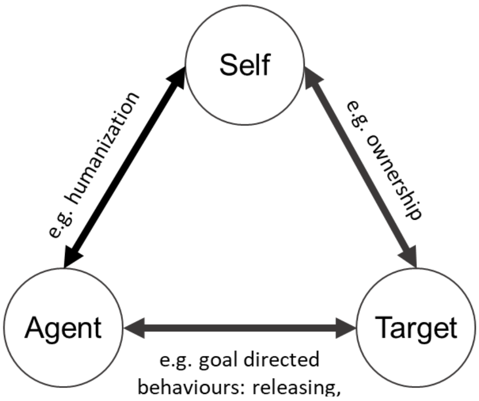


Figure 1-3. A graphic illustration of the mentalizing triangle.

1.3.1 The self-object interaction

I will first expand on the role of self-object relationships as a possible trigger for VPT and mentalising in the triangle model (Figure 1-3. the right edge of the triangle). Some objects become special to us if they are linked to our self-concept. A most prominent example is our names. A name is in nature a set of letters, however as we constantly interact with it, it becomes part of our self-identity and takes priority in attention shifts and memory (Dion, 1983). Recently, studies showed that processing one’s own name directly engages brain systems linked to mentalising. Kampe, Frith, & Frith (2003) found that the mPFC and bilateral temporal poles were activated when someone is calling our names. As these areas also relate closely with the process that we think about other’s intentions (see Frith & Frith, 2006 for a review), this suggests self-relevant information may automatically trigger the process of mentalizing.

Other studies showed that even briefly interacting with particular objects can bias how those objects are processed in our cognitive system. Cunningham (2008) invited participants to the lab and complete a pretending shopping task, where different objects were arbitrarily categorized into 'yours' and 'mine'. After the task, participants were surprisingly given a memory recall task, and results showed that objects classified into the 'mine' category were recalled better. Truong and colleagues further revealed the memory bias effect is modulated how participants interact with the object, that the memory advantage for self-owned object only existed when participants physically interacted with the object, instead of interacting with it on the computer, and such effect was modulated by how closer the objects were places to themselves (Truong, Chapman, Chisholm, Enns, & Handy, 2016).

Recent studies further revealed the self-relevant bias not only lies in the quality of processing, but also manifests in processing speed. Researchers found that neurotypical adults process self-relevant information more rapidly than other information. In one study, participants first learned to associate different labels of geometric shapes to either self, others or neutral terms. With a subsequent label-shape matching task, researchers found that categories previously linked to self were matched faster than categories linked to others (Sui, He, & Humphreys, 2012; Sui, Liu, Mevorach, & Humphreys, 2015). These results indicate a brief interaction would have the potential to change how an object is registered in our mind, which might trigger mentalizing if we find others also encounter those objects.

Another way that the self-object relationship could modulate mentalizing is when we are interacting with valuable objects. Neuroimaging studies showed that our brain is sensitive to the value of targets, that stimuli with high monetary values can trigger the reward circuit which interplays with multiple perceptual systems (Janelle, 2012; Gottlieb, Hayhoe, Hikosaka, & Rangel, 2014). Although until now no study has tested mentalizing with objects of different values, a study has shown that monetary incentive can modulate mentalizing. Cane, Ferguson and Apperly (2017) invited participants to a Director Task, where they needed to follow a virtual director's instruction and take the correct object from a set of shelves. As some of the shelves were opaque, participants sometimes need to think about what can be seen by the director in order to take the object from the common view. They found that when no monetary reward was presented, participants showed no sign of spontaneous VPT. However, in a second block where rewards were given, the objects now have a monetary value, and participants would look to the correct object even before the director give the full instruction. This suggests that valuable object can change our motivation which could further change mentalising performance.

1.3.2 Agent-object interaction

Another factor that might trigger spontaneous mentalizing is the agent-object interaction (Figure 1-3. the bottom edge of the triangle). Almost all paradigms in the ToM research involves interaction between the agent and the target object. In the famous Sally-Anne task and its adapted versions, the goal-directed actions are often embedded in the verbal narrations of the belief stories (Clements & Perner, 1994; Wimmer & Perner, 1983). For example, the experimenter would say “Sally *hid* her marble in the box” (‘hide’ is a goal-directed behaviour), and this also applies to the Duplo task where participants (who are usually children) were invited by the experimenter to play a hiding game with the dummy character (Rubio-Fernández & Geurts, 2013). Such verb descriptions clearly provide clues to participants that the target object was attended to by the agent, i.e. there is an ‘agent-object’ link in this situation. In real life, by interpreting such links, we can infer the agent’s current mental states. Such ability is critical since it provides contents on which spontaneous social interaction can be based. For example, if someone trying to get a box which is placed too high to reach, the correct option would be helping him to get the box, it would be unwise to ignore such a strong ‘reaching’ cue but asking him what he ate for lunch.

Although researchers may not have deliberately designed these interactions, this reflects the tendency that humans always have to interact with their surroundings environment. It seems that we are equipped with specific psychological sensors to detect and interpret such social cues and make sense of the social meaning behind them. Even newborn babies (less than 5-day-old) are able to rudimentarily follow gaze, differentiate biological motions from non-biological motions and demonstrate a preference towards goal-directed actions (Simion, Regolin, & Bulf, 2008; Craighero, Leo, Umiltà, & Simion, 2011). Critically, such spontaneous behaviours were only observed with the social configuration kept intact (Shi, Weng, He & Jiang, 2010; see Palermo & Rhodes, 2007 for a review). The performance on perceiving such cues is found to be related to a variety of social cognitive competences in both children and adults (Frischen, Bayliss, & Tipper, 2007; Gredebäck, Fikke, & Melinder, 2010; Klin, Lin, Gorrindo, Ramsay, & Jonas, 2009; Pavlova, 2012; Sun, Stein, Liu, Ding, & Nie, 2017; Gao, Ye, Shen, & Perry, 2016; He, Guo, Zhai, Shen, & Gao, 2018). It even remains intact when general intelligence is impaired among population with developmental delay (Sparrow, Shinkfield, Day, & Zerman, 1999). However, individuals with autism and schizophrenia who are impaired on mentalizing ability are often reported to be impaired in processing such cues (Blake, Turner, Smoski, Pozdol, & Stone, 2003; Freitag, Konrad, Häberlen, Kleser, von Gontard, Reith, & Krick, 2008; Schultz, Gauthier, Klin, Fulbright, Anderson, Volkmar, & Gore, 2000; Kim, Doop, Blake, & Park, 2005).

Such goal-directed behaviours are also common in ToM research. In the adult non-verbal FBT, agents in the video would usually put the object on the table at the beginning of each trial, rather than being a mere observer (He et al., 2012; Low & Watts, 2013). This incidental detail, however, provides a strong social cue which is likely to implicitly motivate the participants to imagine a relationship between the object and the agent: perhaps the ball belongs to the agent so that the agent would care about the ball location. Similar goal-directed behaviour was also introduced in Senju's anticipatory looking paradigm: the agent would always *reach* the target at the end of each trial (A. Senju et al., 2009; Atsushi Senju, 2013). And this "reach" action provides useful information on the following movement of the agent in the test trial and gives reasons for the participants to track the agent's belief about the ball location.

Although the aims and designs varied across different studies, goal-directed behaviours (either hide, or reach, or simply put on the table) coincidentally, existed in all these studies. This gives rise to the idea that the 'agent-object' interaction is likely to change how mentalizing manifests. Unfortunately, to date, no study has made the agent into a pure observer, so this hypothesis has not yet been tested. Only in one study, the goal-directed behaviour was less obvious to the participants. Using the anticipatory looking paradigm, Schneider et al. (2013) only displayed the agent's '*reaching*' behaviour in filler trials, while in the formal trials the agent became a mere observer. They revealed a weak anticipatory looking effect, as such looking behaviour vanished when participants were exposed to a low cognitive load task (see https://www.youtube.com/watch?v=HMaLIBRwN-Q&feature=youtu.be for a demo). Thus, it is possible to presume that if the agent didn't act in a goal-direct way in the filler trials, anticipatory looking may not exist even in no load condition since participants would have no reason to do so.

Such results above highlight the potential role of 'agent-object' interaction in triggering the mentalizing process, by pointing out the inclusion of 'goal-directed' behaviours in previous studies. However, the 'agent-object' interaction may not be limited by 'goal-directed' behaviours. We consider the reason that goal-directed behaviours can promote mentalizing is because they suggest a change in the 'agent-object' relationship. Thus, it's plausible to presume that any information which can suggest the link between the agent and the object should have similar effect. For example, specific ‘agent-object’ relationship might also be delivered by explicit instructions, or background stories, such as a narrative from the agent. The critical point is such information provides participants a reason or a goal to engage in mentalizing, thus they are more likely to do so.

1.3.3 Self-agent interaction: the role of dehumanization

A prerequisite for mentalizing is that we detect an agent who has a mind. This may sound intuitive, however previous research suggests that we do not attribute equal humanness to everyone. Such inequality shapes different social relationships between self and others, and more importantly, it alters the ‘self-agent’ interaction could therefore alter the propensity to engage in mentalising (Figure 1-3. the left edge of the triangle).

The process of considering other people to have less than full mental capacities is named dehumanization. According to Haslam (2009), dehumanization process is due to a category errors made by our cognitive system, thus we perceive some others more like the way we see animals or objects. Dehumanization can also take a more subtle and everyday form, that we perceive some of us in an abstract way and attribute less human essence to them, such a phenomenon is called infrahumanization. Haslam claimed that both dehumanization and infrahumanization can act on two aspects of human essence: a person’s unique human features (e.g. moral judgments) or human nature (e.g. emotionality or creativity). Denying a person’s unique human feature results in seeing them as more animal-like; while denying their human nature lead to compare others to automata or machine (Haslam, 2006).

Leyens and colleagues (2000) described how individuals infrahumanize others in terms of unique human features. They categorized human emotions into primary emotions, which shared by other animals and humans, and secondary emotions that uniquely belong to humans. In a critical study, they found that when individuals were randomly assigned to a group, they attributed fewer secondary emotions to out-group members, suggesting a denial of the out-group individual’s unique human features. Buckels & Trapnell (2013) also found that in the intergroup context, people were inclined to implicitly associate out-group members with animals, especially when they were evoked with disgusting emotions. In parallel, people spontaneously deny some group’s human nature and liken them with automata or machine. Such studies demonstrated that even within weak intergroup settings, people tend to treat ingroup and outgroup others differently.

Dehumanization then impacts on other cognitive process and bring different behavioural results. Fincher, Tetlock and Morris (2017) found that with dehumanized individuals, our face recognition shifted from configural processing to processing a collection of features. Such a change in processing manner may disrupt the social information delivered by other people's face or body, such as changing the threshold for us to perceive other's goal-directed actions.

Dehumanization can also have a direct impact on mentalizing, as inferring mental states from dehumanized targets is disrupted (Harris & Fiske, 2011). Neuroimaging studies even showed that participants failed to recruit the brain networks used for everyday social cognitive processes when dehumanizing others (Harris & Fiske, 2006). Dehumanized individuals are regarded as more insensitive to pain and evoked less empathetic responses (Haslam, 2006; Haslam, Bain, Lee, Douge, & Bastian, 2005; Haslam, Loughnan, Kashima, & Bain, 2008). A related effect is that people expressed less willingness help to dehumanized groups (Cuddy, Fiske, & Glick, 2007) and intergroup helping is indeed found to be reduced in times of need (Cuddy, Rock, & Norton, 2007). These evidences accumulated and suggest at least some forms of social cognition is modulated by how the agent is perceived by us as a human, thus, the 'self-agent' interaction may potentially change our propensity to engage in mentalizing.

1.3.4 Mentalizing in the mentalizing triangle

When do people engage in mentalizing? Why sometimes we fail to process the mental content of some others? To answer these questions, I summarized the previous findings on mentalizing and integrated them in a triangle model composed of self, agent(s) and object(s). The previous sections reviewed evidence of how self-object, agent-object and self-agent relationship might change the propensity to engage in mentalizing. I hypothesise that activation of any of the three sides of the mentalising triangle can trigger the mentalising process. First, factors linking an object to the self can trigger mentalising about who else can see that object (self-object interaction). Second, social cues like seeing goal directed actions allow us to link an agent to an object and trigger mentalising about what that person knows about the object (agent-object interaction). Third, when we humanise another person and consider them a full person, we can engage in mentalising about what they know (self-agent interaction). In the following chapters, I systematically manipulated the agent-object interaction and the self-agent interaction, and tested whether mentalizing changes according to these variations.

1.4 Research methods on investigating mentalizing

In paragraphs above I summarized the debates around evidence on VPT and ToM. A key issue is whether an implicit mentalizing system exists. One possible reason for these debates is that measurements in these tasks usually only rely on a single behavioural index, such as reaction time, or anticipatory looking behaviour. For example, the majority of research on implicit mentalizing in adults uses RT as the main behavioural index. However, RT is easily influenced by a number of social and non-social factors. As mentioned above, Phillips et al (2015) challenged Kovács et al. (2010) findings on implicit mentalizing by reporting that their RT difference may due to different time points of the 'keyhit' events before participants made a belief-relevant response. Thus, solely rely on RT might lead to misinterpretations of the results. Another limitation comes from the standard experimental setting. In the majority of the mentalizing tasks, researchers used static figures as stimuli. However, real life social interactions are composed of moving agents and immersive feelings. Therefore, participants may lose interest in engaging interactions with static figures, which might be the cause of an absence of mentalizing in some studies.

To import new measurements in mentalizing, recently a novel study combined the classic signal detection theory (SDT) with a level-1 VPT task (Seow & Fleming, 2019). Signal detection theory examines how people distinguish valid signal from the distribution of noise. It divided the human decision process into two components: perceptual sensitivity, which is represented as dprime (*d'*) and the decision criteria (*c*). These two indices are relatively independent of each other and can be influenced by distinct factors. For example, participants' sensitivity to detect near-threshold visual stimuli can be driven by the contrast of the stimuli, but their tendency to report seeing a signal might be more easily related to the reward/punishment attached to different answer types.

Taking advantage of the SDT, Seow & Fleming investigated whether observing near-threshold stimuli with another person can change our perception in the context of the Samson dots task (see Figure 1-2A, Samson et al., 2015). For each trial, participants were instructed to report if a near-threshold Gabor patch was presented on the wall in front of or behind a virtual agent. Results showed that when the agent is able to see the target (oriented towards the target and the vision is not blindfolded), participants demonstrated lower perceptual threshold to detect the Gabor and had a more liberal decision criteria, suggesting they tended to believe they saw the target. This study offers a new method to investigate how other's mental contents influence our cognitive processes, by providing more systematic decoding of the decision-making process, thus has the potential to address the debate brought by the RT method. In Chapter 2 and 3, we will combine this measurement with a false belief task. In addition, we implemented a Bayesian Hierarchical Model analysis on these datasets. These improvements will provide more detailed and reliable information of how other's beliefs influence participants' perception.

Studies also suggested that more ecological and diverse methods might help to uncover the multi-faceted mentalizing processes. Freundlieb et al. (2016, 2017, 2018) invited real people as the confederates in their VPT tasks, and provide solid evidence on spontaneous VPT effect in their settings. However, including real human beings may bring other confounding factors such as race, appearance and unexpected social communications, and is not suitable for cross-cultural studies and meta-analysis. To balance the needs of standard control and ecological validity, using videos with a real person or implementing studies in Virtual Reality (VR) may offer a good solution. In Chapter 4 and 5, we designed a social mental rotation task with exhibition of moving agents in an immersive VR environment. We believe such effort could better simulate real life mentalizing scenarios, thus give larger explanatory power to our results.

1.5 Conclusion

In this chapter, I briefly introduced the concepts of two core components in mentalizing: theory-of-mind and visual perspective taking, and summarized results and indications from current research. I described the two-system view on mentalizing, and pointed out its limitation in guiding research in real-life settings. Based on that, I put forward a new model based on a mentalizing triangle, which is composed of the interacting self, agent and object. Crucially, I argue that the dynamic interactions within the triangle matters for promoting mentalizing. Thus, when self, agent and the object are not linked with each other, it is unlikely for people to engage in mentalizing. Meanwhile, I pointed out the necessity of importing new approaches to explore problems in this field.

In the next chapters, I will present four experimental studies. The first study targets ToM in the minimal social setting, i.e. there is no interaction between any dyadic relationships in the triangle. Based on that, the second study added the 'agent-object' interaction in the mentalizing process, aiming to observe a stronger spontaneous mentalizing effect. In Chapter 4 & 5, I investigated the ‘self-agent’ interaction in a VPT task, specifically, how would people select which perspective to take when exposed to conflicting perspectives. Importantly, I explored the relevant neural mechanisms underlying such selection and discussed how the humanization process interacts with VPT. In the last chapter, I summarized all the findings in this thesis and discuss current research limitations, as well as how these current results relate to previous studies and benefit future research.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states?. *Psychological review, 116*(4), 953. https://doi.org/10.1037/a0016923

Baron-Cohen, S. E., Tager-Flusberg, H. E., & Cohen, D. J. (2000). *Understanding other minds: Perspectives from developmental cognitive neuroscience*. Oxford University Press.

Baron-cohen, S., Leslie, A., & Frith, U. (1985). The autistic child have a “theory of mind”? *Cognitive Development*, *21*, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes” Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Beck, A. A., Rossion, B., & Samson, D. (2018). An objective neural signature of rapid perspective taking. *Social Cognitive and Affective Neuroscience*, *13*(1), 72–79. https://doi.org/10.1093/scan/nsx135

Bukowski, H. (2018). The Neural Correlates of Visual Perspective Taking: a Critical Review. *Current Behavioral Neuroscience Reports*, *5*(3), 189–197. https://doi.org/10.1007/s40473-018-0157-6

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind and Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Cane, J. E., Ferguson, H. J., & Apperly, I. A. (2017). Using perspective to resolve reference: The impact of cognitive load and motivation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(4), 591–610. https://doi.org/10.1037/xlm0000345

Capozzi, F., Cavallo, A., Furlanetto, T., & Becchio, C. (2014). Altercentric intrusions from multiple perspectives: Beyond dyads. *PLoS ONE*, *9*(12), 1–14. https://doi.org/10.1371/journal.pone.0114210

Christensen, W., & Michael, J. (2016). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, *40*, 48–64. https://doi.org/10.1016/j.newideapsych.2015.01.003

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive development, 9*(4), 377-395.

Cox, S., Chandler, C., Simpson, A., & Riggs, K. (2016). The effect of alcohol dependence on automatic visuo-spatial perspective taking. *Drug and Alcohol Dependence, 166,* 21–25. https://doi.org/10.1016/j.drugalcdep.2016.06.007

Dumontheil, I., Küster, O., Apperly, I. A., & Blakemore, S. J. (2010). Taking perspective into account in a communicative task. *NeuroImage*, *52*(4), 1574–1583. https://doi.org/10.1016/j.neuroimage.2010.05.056

Elekes, F., Varga, M., & Király, I. (2016). Evidence for spontaneous level-2 perspective taking in adults. *Consciousness and Cognition*, *41*, 93–103. https://doi.org/10.1016/j.concog.2016.02.010

Flavell, J. H. (1977). The development of knowledge about visual perception. In *Nebraska symposium on motivation*. University of Nebraska Press.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of “theory of mind” in story comprehension. *Cognition*, *57*(2), 109–128. https://doi.org/10.1016/0010-0277(95)00692-R

Freundlieb, M., Kovács, Á. M., & Sebanz, N. (2016). When do humans spontaneously adopt another’s visuospatial perspective? *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 401–412. https://doi.org/10.1037/xhp0000153

Freundlieb, M., Kovács, Á. M., & Sebanz, N. (2018). Reading your mind while you are reading—evidence for spontaneous visuospatial perspective taking during a semantic categorization task. *Psychological science*, *29*(4), 614-622.

Freundlieb, M., Sebanz, N., & Kovács, Á. M. (2017). Out of Your Sight , Out of My Mind : Knowledge About Another Person ’ s Visual Access Modulates Spontaneous Visuospatial Perspective-Taking. *Journal of Experimental Psychology : Human Perception and Performance*, *Online*(6), 1065–1072. https://doi.org/10.1037/xhp0000379

Frischen, A., Bayliss, A., & Tipper, S. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694–724.

https://doi.org/10.1037/0033-2909.133.4.694.Gaze

Frith, C. D., & Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron*, *50*(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Furlanetto, T., Becchio, C., Samson, D., & Apperly, I. (2016). Altercentric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(2), 158.

Gao, Z., Ye, T., Shen, M., & Perry, A. (2016). Working memory capacity of biological movements predicts empathy traits. *Psychonomic bulletin & review*, *23*(2), 468-475.

Gredebäck, G., Fikke, L., & Melinder, A. (2010). The development of joint visual attention: A longitudinal study of gaze following during interactions with mothers and strangers. *Developmental Science*, *13*(6), 839–848. https://doi.org/10.1111/j.1467-7687.2009.00945.x

He, J., Guo, D., Zhai, S., Shen, M., & Gao, Z. (2018). *Development of Social Working Memory in Preschoolers and Its Relation to Theory of Mind*. *00*(0), 1–14. https://doi.org/10.1111/cdev.13025

He, Z., Bolz, M., & Baillargeon, R. (2012). 2.5-Year-Olds Succeed At a Verbal Anticipatory-Looking False-Belief Task. *British Journal of Developmental Psychology*, *30*(1), 14–29. https://doi.org/10.1111/j.2044-835X.2011.02070.x

Kampis, D., Parise, E., Csibra, G., & Kovács, Á. M. (2015). Neural signatures for sustaining object representations attributed to others in preverbal human infants. *Proceedings. Biological Sciences / The Royal Society*, *282*(1819), 20151683-. https://doi.org/10.1098/rspb.2015.1683

Klin, A., Lin, D. J., Gorrindo, P., Ramsay, G., & Jonas, W. (2009). Two-year-olds with autosm orient to nonsocial contigencies rather than biological motion. *Nature*, *459*(7244), 257–261. https://doi.org/10.1038/nature07868.Two-year-olds

Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, *330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Kristine H. Onishi and Renée Baillargeon. (2005). *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621.Do

Low, J., & Perner, J. (2012). Implicit and explicit theory of MD: State of the art. *British Journal of Developmental Psychology*, *30*(1), 1–13. https://doi.org/10.1111/j.2044-835X.2011.02074.x

Low, J., & Watts, J. (2013). Attributing False Beliefs About Object Identity Reveals a Signature Blind Spot in Humans' Efficient Mind-Reading System. *Psychological Science*, *24*(3), 305–311. https://doi.org/10.1177/0956797612451469

Lyons, M., Caldwell, T., & Shultz, S. (2010). Mind-reading and manipulation - Is Machiavellianism related to theory of mind? *Journal of Evolutionary Psychology*, *8*(3), 261–274. https://doi.org/10.1556/JEP.8.2010.3.7

MacDorman, K. F., Srinivas, P., & Patel, H. (2013). The uncanny valley does not interfere with level 1 visual perspective taking. *Computers in human behavior*, *29*(4), 1671-1685.

Moll, H., & Meltzoff, A. N. (2011). How Does It Look? Level 2 Perspective-Taking at 36Months of Age. *Child Development*, *82*(2), 661–673. https://doi.org/10.1111/j.1467-8624.2010.01571.x

Moll, H., Meltzoff, A. N., Merzsch, K., & Tomasello, M. (2013). Taking versus confronting visual perspectives in preschool children. *Developmental Psychology*, *49*(4), 646–654. https://doi.org/10.1037/a0028633

Pavlova, M. A. (2012). Biological motion processing as a hallmark of social cognition. *Cerebral Cortex*, *22*(5), 981–995. https://doi.org/10.1093/cercor/bhr156

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, *16*(10), 1706–1716. https://doi.org/10.1162/0898929042947900

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences*, *16*(10), 519–525. https://doi.org/10.1016/j.tics.2012.08.004

Ramsey, R., Hansen, P., Apperly, I., & Samson, D. (2013). Seeing It My Way or Your Way: Frontoparietal Brain Areas Sustain Viewpoint-independent Perspective Selection

Processes. *Journal of Cognitive Neuroscience*, *25*(5), 670–684. https://doi.org/10.1162/jocn_a_00345

Rene´e Baillargeon, R. M. S. and Z. H. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Repacholi, B., & Slaughter, V. (Eds.). (2004). *Individual differences in theory of mind: Implications for typical and atypical development*. Psychology Press.

Rubio-Fernández, P., & Geurts, B. (2013). How to Pass the False-Belief Task Before Your Fourth Birthday. *Psychological Science*, *24*(1), 27–33. https://doi.org/10.1177/0956797612447819

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & 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. https://doi.org/10.1037/a0018729

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain*, *128*(5), 1102–1111. https://doi.org/10.1093/brain/awh464

Sartori, L., Becchio, C., & Castiello, U. (2011). Cues to intention: The role of movement information. *Cognition*, *119*(2), 242–252. https://doi.org/10.1016/j.cognition.2011.01.014

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind.” *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Schwarzkopf, S., Schilbach, L., Vogeley, K., & Timmermans, B. (2012). Automatic and Intentional Level 1 Perspective-Taking in Adults with High-Functioning Autism. *Proceedings of KogWis 2012*, 145.

Schneider, D., Nott, Z. E., & Dux, P. E. (2014). Task instructions and implicit theory of mind. *Cognition*, *133*(1), 43–47. https://doi.org/10.1016/j.cognition.2014.05.016

Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition*, *129*(2), 410–417.

https://doi.org/10.1016/j.cognition.2013.08.004

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition*, *162*, 27–31. https://doi.org/10.1016/j.cognition.2017.01.018

Schurz, M., Kronbichler, M., Weissengruber, S., Surtees, A., Samson, D., & Perner, J. (2015). Clarifying the role of theory of mind areas during visual perspective taking: Issues of spontaneity and domain-specificity. *NeuroImage*, *117*, 386–396. https://doi.org/10.1016/j.neuroimage.2015.04.031

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind Eyes: An Absence of Spontaneous Theory of Mind in Asperger Syndrome. *Science*, *325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Senju, A. (2012). Spontaneous theory of mind and its absence in autism spectrum disorders. *The Neuroscientist*, *18*(2), 108-113.

Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, *105*(2), 809–813. https://doi.org/10.1073/pnas.0707021105

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Sparrow, W. A., Shinkfield, A. J., Day, R. H., & Zerman, L. (1999). Visual perception of human activity and gender in biological-motion displays by individuals with mental retardation. *American journal on mental retardation*, *104*(3), 215-226.

Sun, Y., Stein, T., Liu, W., Ding, X., & Nie, Q. Y. (2017). Biphasic attentional orienting triggered by invisible social signals. *Cognition*, *168*(July), 129–139. https://doi.org/10.1016/j.cognition.2017.06.020

Todd, A. R., Cameron, C. D., & Simpson, A. J. (2017). Dissociating processes underlying level-1 visual perspective taking in adults. *Cognition*, *159*, 97–101. https://doi.org/10.1016/j.cognition.2016.11.010

Troje, N. F. (2013). What Is Biological Motion? Definition, Stimuli, and Paradigms. *Social*

Perception, 13–36. https://doi.org/10.7551/mitpress/9780262019279.003.0002

Ward, E., Ganis, G., & Bach, P. (2019). Spontaneous Vicarious Perception of the Content of Another's Visual Perspective. *Current Biology*, *29*(5), 874-880.e4. https://doi.org/10.1016/j.cub.2019.01.046

Whiten, A. (1997). The machiavellian mindreader. *Machiavellian intelligence II: Extensions and evaluations*, 144-173.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Chapter 2. Spontaneous attribution of false beliefs in adults examined using a signal detection approach

This chapter has been published on Quarterly Journal of Experimental Psychology as: Ye, T., Fleming, S. M., & Hamilton, A. F. (2019). Spontaneous attribution of false beliefs in adults examined using a signal detection approach. *Quarterly Journal of Experimental Psychology*, 1747021819884677.

2.1 Abstract

Understanding other people have beliefs different from ours or different from reality is critical to social interaction. Previous studies suggest that healthy adults possess an implicit mentalising system, but alternative explanations for data from reaction time false belief tasks have also been given. In this study, we combined signal detection theory (SDT) with a false belief task. Since the application of SDT allows us to separate perceptual sensitivity from criteria, we are able to investigate how another person's beliefs change the participant's perception of near-threshold stimuli. Participants (n=55) watched four different videos in which an actor saw (or didn't see) a Gabor cube hidden (or not hidden) behind an occluder. At the end of each video, the occluder vanished revealing a cube either with or without Gabor pattern, and participants needed to report whether they saw the Gabor pattern or not. A pre-registered analysis with classical statistics weakly suggests an effect of the actor's belief on participant's perceptions. An exploratory Bayesian analysis supports the idea that when the actor believed the cube was present, participants made slower and more liberal judgements. Though these data are not definitive, these current results indicate the value of new measures for understanding implicit false belief processing.

Keywords: theory-of-mind, false belief task, signal detection theory, Bayesian hierarchical model

2.2 Introduction

In everyday social interaction, people often need to track other's mental states (eg. beliefs) either to understand their behaviours or generate social responses (Carruthers, 2015; Frith, 1999; Frith & Frith, 2005; Koster-Hale & Saxe, 2013; Samson, 2013). Recently, attention has focused on the question of the cognitive mechanisms needed to engage in mentalising, and it has been proposed that people may be able to engage in implicit mentalising processes without the need for language or executive processing. Such implicit mentalising would make mindreading easier and more efficient (Apperly, 2009; Onishi & Baillargeon, 2005; Low & Perner, 2012; Baillargeon, Scott & He, 2010; Clements & Perner, 1994). However, the question of whether this mechanism really exists is still hotly debated (Apperly, 2009; Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006; Back & Apperly, 2010; Low & Perner, 2012; Heyes, 2014; Leslie, Friedman & German, 2004).

The strongest evidence for spontaneous and implicit processing of other people's mental states comes from studies of altercentric intrusion effects, that is, cases where the mental state of another person impacts on the ability to do an individual task. Such effects were first demonstrated in the case of visual perspective taking (VPT) (Samson et al., 2010). In this study, participants see a room with an avatar in the centre looking towards one wall, and between one and three red discs on the walls either in front of or behind the avatar. They must report how many discs they can see, and results showed that when participant's and the avatar's perspective are incongruent (e.g. when a participant can see three discs but the avatar can only see two), participants responded slower and make more errors. Since participants were not required to take the avatar's perspective, such altercentric intrusion suggests that participants spontaneously considered other's perspective. However, the interpretation of these results remains controversial, with similar effects found for arrows (Santiesteban, Catmur, Hopkins, & Bird, 2014), implying a non-social explanation of the effect. Conversely, if participants see an avatar who wears opaque or transparent goggles, the altercentric intrusion is found only for the transparent goggles which implies a genuine mentalising explanation of the effect (Furlanetto, Becchio, Samson, Apperly, 2015; but see Conway, Lee, Ojaghi, Catmur & Bird, 2017).

A recent study builds on these results by showing that altercentric intrusion can also improve performance on a perceptual task. Seow and Fleming (2019) combined a level-1VPT task with a signal detection paradigm in which participants must detect a near-threshold Gabor pattern on a grey disc, presented on the wall of Samson's room. On some trials, the avatar could also see the disc and other times he could not, either because he is facing the wrong way or is wearing a blindfold. Results showed that when participants knew that the avatar could see the disc, their perceptual sensitivities (for judgements from the participant's perspective) were enhanced as the *d'* in this condition were higher than in which the avatar could not see (Seow & Fleming, 2019). This study indicates that processing another person's level-1 visual perspective can occur spontaneously and can influence low-level perceptual processes.

However, understanding what another person can see is only a starting point for full mentalising, a more important building block is how we interpret other's beliefs. Recent studies suggested that people could implicitly predict other's behaviours based on their current beliefs, even on those conflicting with reality. A widely-used task here is the false belief task, where a protagonist is sometimes misled to believe that the target is hidden in a wrong location (or not), and participants at some points are asked to predict where the protagonist would look for the target (Wimmer & Perner 1983). Recent studies adapted this task into a non-verbal version and found participants tended to look to where the protagonist believed the target was hidden (Southgate, Senju & Csibra, 2007; Senju, Southgate, White & Frith, 2009). Later Schneider and colleagues (Schneider, Lam, Bayliss & Dux, 2012; Schneider, Nott & Dux, 2014) replicated these results and claimed that anticipatory looking could be observed even after participants have performed the task for 1hr. Neuroimaging studies further suggested that implicit behaviours share large overlap in neural substrates with those involved in explicit mentalising, but the former might be more sensitive to belief contents (Naughtin, Horne, Schneider, Venini, York & Dux, 2017; Kovács, Kühn, Gergely, Csibra, Brass, 2014; Schneider, Slaughter, Becker & Dux, 2014).

Nonetheless, there are also studies questioned if anticipatory looking can be taken as a robust behavioural sign for implicit mentalising as such behaviours were found to be influenced by task instructions and cognitive load (Cane, Ferguson & Apperly, 2017; Schneider et al., 2012b; Schneider et al., 2014a). An alternative approach was taken by Kovács and colleagues (2010), who designed a novel reaction-time based false belief task, where implicit mentalising was examined when participants were doing a belief-irrelevant task. In this study, participants were required to judge as fast as possible if a ball was present or not when an occluder vanished. Before they made this judgement, they saw a video clip in which an agent saw (or didn't see) a ball hiden behind the occluder (or not). This design resulted in the agent having a true or false belief about the ball location, which was not relevant to the participant's task of deciding if the ball was present. Despite this, participants did respond faster if the agent believed the ball was behind the occluder, which was taken as evidence for spontaneous, implicit theory of mind.

Again, there has been controversy surrounding this finding. A recent paper from Phillips et al. (2015) presented 13 experiments based on Kovács' method and claimed that the reaction time pattern in Kovács' study cannot only be interpreted in terms of belief-relevant factors. Instead, these researchers concluded that the results from Kovács' study were due to refractory movements rather than belief attribution. With these ambiguous results, it remains hard to draw a strong conclusion about whether adults spontaneously and implicitly ascribe beliefs to others or not. One key reason for these confusing results is that there are many factors influencing performance on reaction time tasks, including interference from previous actions and the physical features of the stimuli (Herman & Kantowitz, 1970; Niemi & Näätänen, 1981). Therefore, it may be useful to find alternative tasks which do not rely solely on reaction time measures. In this respect, Seow and Fleming's study offers a new approach to investigate belief ascription in the context of signal detection theory (SDT).

SDT provides a framework for decomposing perceptual performance into two statistics: *d'*, the sensitivity of the system to the occurrence of the signal in signal-to-noise ratio units, and an overall bias to report signal presence, modelled as the criterion, *c*. This allows us to separate the perceptual decision-making criteria from individual differences in perceptual ability or reaction time (Rouder et al., 2005; Stanislaw & Todorov, 1999) . Seow's study revealed that knowing another person can see the same target as us can increase our perceptual sensitivity and change our decision criteria. In the present paper, we further predict that knowing the other person has seen something (or not) could also change perceptual sensitivity or decision criteria. Importantly, when either participants or an agent believed a target is present, participants should have a higher expectation that the target is present and should tend to report that they see the target. Hence, their perceptual sensitivities may increase while criteria may drop as participant's decisions becoming more liberal. In this way, SDT may provide a more sensitive measure of implicit mentalising.

In this current study, before I test whether the relationships within the triangle model could impact on spontaneous mentalizing behaviours, I first would like to seek evidence fot spontaneous mentalizing behaviour when no interaction exists among the self, agent and object relationships. I hypothesized that when there is no interaction among the triplets within the triangle model, there might be no or very weak spontaneous mentalizing behaviour. I thus applied a psychophysical approach that allows calculation of SDT measures to a version of Kovács' non-verbal false belief task. That is, I asked participants to detect a target feature (Gabor pattern) in contexts where another person either does or does not believe that the target feature is present. I then tested if the altercentric intrusion from another person's belief can influence participant's perceptual sensitivity (*d'*) or decision criterion (c). I termed his paradigm a feature-detection false belief task. The methods and analysis of this study were pre-registered with the Open Science Framework (please see link https://osf.io/wxy2p/).

2.3 Method

Participants

Our target sample size was 40 participants. To achieve this, 66 participants were recruited to the experiment from the UCL-ICN participant database and were paid at a rate of £7.5 per hour. The study is under the ethical approval from the UCL Research Ethics Committee and conformed to the 1964 Declaration of Helsinki. Following the pre-registered data exclusion criteria, 26 participants were excluded from the original dataset, leaving a final sample of n=40 (16 males, age 24.6 ± 3.5).

The exclusion criteria are:

1. A participant scores 33 or above on the Autistic-Spectrum Quotient (AQ) test, which is suggested as a cut-off point between typical performance and autism (Baron-Cohen, Wheelwright, Skinner, Martin & Clubley, 2001);
2. A participant performs below 80% accuracy on the attention check trials.
3. The participant's overall accuracy on the belief tasks is below 55% or above 95%. Lower or higher accuracy will lead to extreme and unstable estimation of SDT measures such as *d'*.
4. A participant has 3 or more than 3 blocks in which biased responses are given. Chi-square test was used to examine whether participants' yes/no responses were significantly different from 50% yes and 50% no. With 24 trials in a block and a 0.05 significance level, if the number of 'Yes' responses was greater than 16 or less than 8, then participant's response profile for this block was significantly different from 50%

‘Yes’ and we considered this to be a biased block (Corder & Foreman, 2014). Participants with more than 3 biased blocks over the study were excluded.

Stimuli & Materials

Our stimuli were a set of 4 video clips in which a cube with a high-contrast grating on one side moves behind a barrier and then leaves the scene (or not) while an observer watches the cube move (or not) based on the logic of Kovács (2010) (Figure 2-1). We use the labels P (participant) and A (actor) to index who believes the moving cube is present (+) or absent (-). The event sequences were as follows:

1) P-A-: The cube moves behind the occluder and then moves out, all under the actor’s gaze. The cube is last seen by the participant at 12.5s. Then the actor turns away from the table and starts to turn towards the table at 15s. The occluder disappears at 17s. In this condition, both the participant and actor hold a ‘target absent’ belief;

2) P-A+: The cube moves behind the occluder. Then the actor turns away from the table. While she is not looking, the cube moves out of the occluder and falls off the table (and out of the screen). The cube is last seen by the participant at 12.5s. Then the actor starts to turn towards the table at 15s. The occluder disappears at 17s. In this condition, the participant believes the target is absent but the actor has a false ‘target present’ belief;

3) P+A-: The cube moves behind the occluder and then moves out and falls off the table (and out of the screen), all under the actor’s gaze. Then the actor turns away from the table. While she is not looking, a same cube moves in behind the occluder. The cube is last seen by the participant at 12.5s. Then the actor starts to turn towards the table at 15s. The occluder disappears at 17s. In this condition, the participant believes the target is present but the actor has a false ‘target absent’ belief;

4) P+A+: The cube moves behind the occluder. Then it moves out of the occluder to the right edge of the table and moves back behind the occluder. These all take place under the actor’s gaze. The cube is last seen by the participant at 12.5s. Then the actor turns away from the table and starts to turn towards the table at 15s. The occluder disappears at 17s. In this condition, both the participant and the actor hold a ‘target present’ belief.

To create these video clips, we recorded the cube movements and the actor movements separately then superimposed them together using Adobe Premiere (Adobe, USA). In each video clip, the 3D environment and the cube's movement trajectories were generated in Vizard 5.7 (WorldViz, USA). All the movements took place against a dark grey [51, 51, 51] background. Each video clip involved a wooden table, a white occluder and a grey [128, 128, 128] cube with Gabor pattern on its front side. The Gabor stimuli on the moving cube contain sinusoidal gratings (contrast 0.25, spatial frequency of 6 cycles per degree and orientation 30 degrees.), superimposed with 10% white noise modulated by a Gaussian envelope.

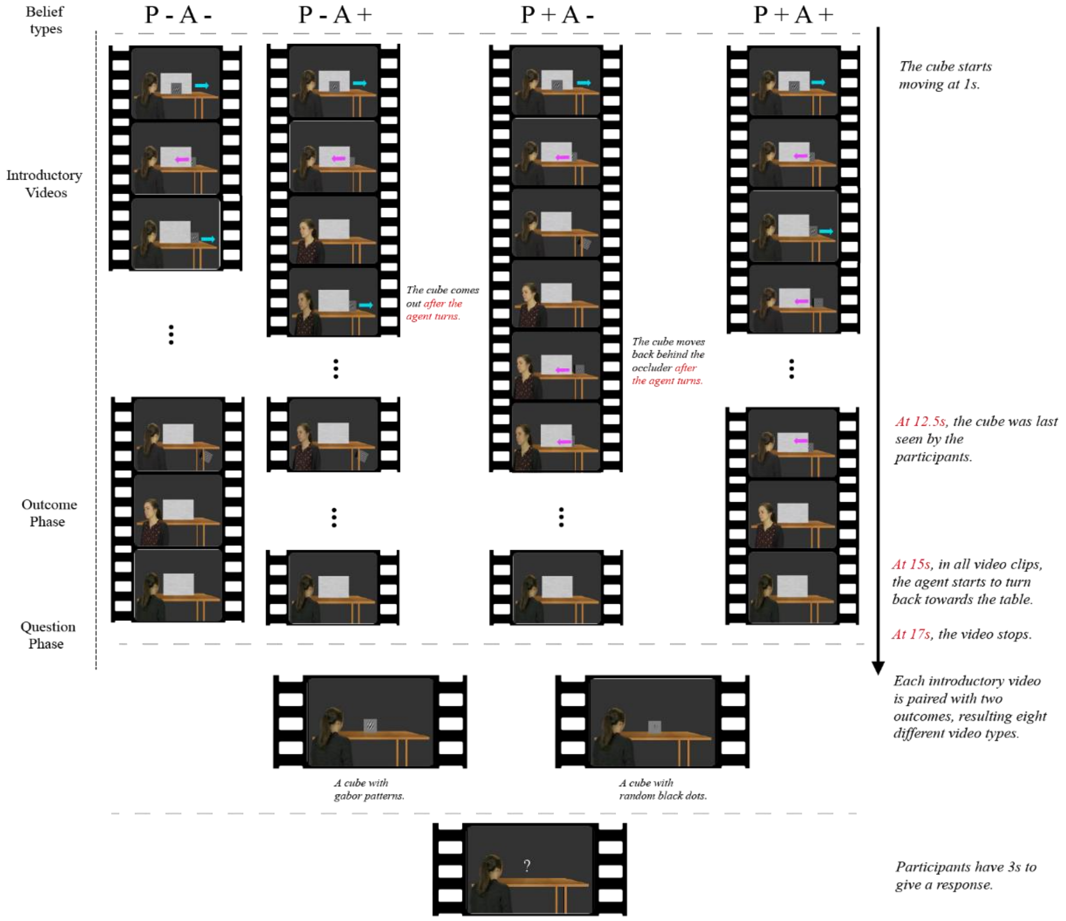

Belief types
P - A -
P - A +
P + A -
P + A +
Introductory Videos
Outcome Phase
Question Phase
The cube starts moving at 1s.
The cube comes out after the agent turns.
The cube moves back behind the occluder after the agent turns.
At 12.5s, the cube was last seen by the participants.
At 15s, in all video clips, the agent starts to turn back towards the table.
At 17s, the video stops.
Each introductory video is paired with two outcomes, resulting eight different video types.
A cube with gabor patterns.
A cube with random black dots.
?
Participants have 3s to give a response.

Figure 2-1. A schematic illustration of all conditions in the belief task. The four different starting events are illustrated as four downward film strips. Pink and blue arrows were not present in the videos but are included here to illustrate ball motion. All four clips ended with one of two outcome phases and then a single question screen.

After the moving cube stimuli were generated, movements of a single actor were recorded in front of a blue background. The actor was required to stand still, look or turn following the director's orders, avoiding excessive movements or facial expressions. To achieve a close match between the actor’s and the cube’s movements for each condition, while filming the actor’s movements, the corresponding cube’s video was played aside to allow the director to give voice commands at the right time. Then the matched actor’s and cube’s videos were merged in Adobe Premiere Pro CC 2017, using Chromakey to remove the blue background. The position and sizes of the actor were carefully matched across conditions and each video was cut to 17 seconds duration and saved without sound and with a resolution of 1024*768 pixels. The final frame in each video was the one before the occluder disappeared. We also saved a test frame from each video, which was the frame immediately after the occluder vanished and depicted the actor facing an empty table. These four pictures were then appended seamlessly at the end of their parent videos in the experimental script to provide a background image for the outcome phase and response phase in the task.

In the experimental trials, a lower-contrast test cube was imposed on these background images, placed in the centre of the table. The test cube was the same grey colour as the moving cube but has either the Gabor pattern or a white noise pattern on its front side. The same background image was used during the thresholding task, where we determined the appropriate threshold for each participant to be able to detect the Gabor pattern (see below). This ensures that the physical environment for Gabor feature detection is identical throughout the experiment.

To monitor the participant’s attention throughout the task, 5 ‘attention check’ videos were created. These video clips were made by editing the belief videos to stop early, before the occluder vanishes. A blue question box appears immediately after the attention video stops, asking a question either about the cube’s current location (on/off the table) or about the actor’s orientation (looking towards / away from the table), but never about the actor’s mental states. In summary, the final stimuli include 4 belief videos matched to 4 background pictures and 5 attention check videos. All the stimuli are presented by Cogent 2000 and Cogent Graphics (http://www.vislab.ucl.ac.uk/cogent.php) in MATLAB (The MathWorks, Natick, MA).

Procedure

When each participant arrived for the study, they read an information sheet about the study and then signed a consent form to take part. Instructions were given verbally by the experimenter, then the participant completed two computer-based tasks: threshold testing and the belief task. Finally, they completed the AQ (Baron-Cohen et al. 2001) and a post-experiment questionnaire. A detailed description of each task is given below.

Threshold testing

This phase of the study aims to measure each participant’s Gabor detection threshold for use in the second part of the study. Participants were asked to detect Gabor patterns of varying contrast in a 2 interval forced choice task. The Gabors were 73 by 73 pixels (210 by 210 mm on the screen and viewed from approx. 60cm), with contrast varying from 0 to 1. As mentioned previously, test stimuli were presented on a grey cube in a scene drawn from the last frame of the video clips.

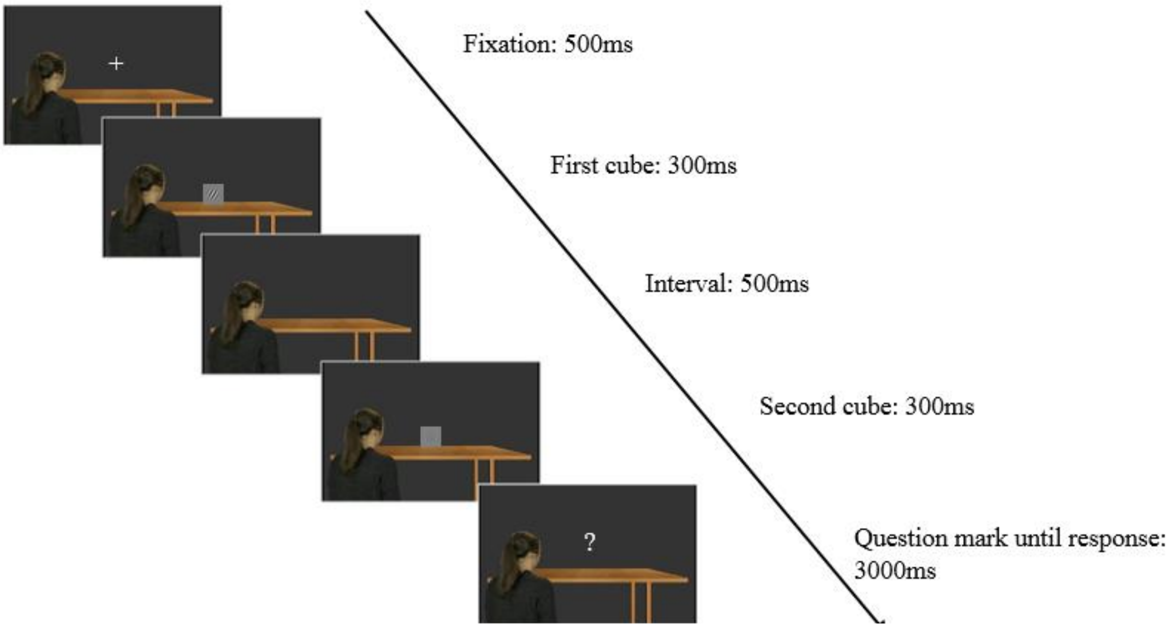


Figure 2-2. Procedure for the threshold testing stage.

Participants were requested to detect which cube has the Gabor pattern, the first or the second. There were 10 practice trials before the formal threshold testing task, and participants were informed that the Gabor pattern would be difficult to detect during the formal session, so that they need to concentrate. In cases where they felt unable to identify the Gabor, they were instructed to make a best guess. There were 120 trials in the threshold testing session.

Each trial started with the onset of the 3D room picture, and then a white fixation appeared followed by the first grey cube, positioned in the middle on the table with a size of ~2 degrees of visual angle. The first cube appeared for 300ms then disappeared, after a short interval (500ms), the second cube appeared for another 300ms (Seow & Fleming, 2019). On half of the trials, the Gabor pattern was on the first cube and for the other half on the second, subtending ~1.84 degrees of visual angle to the centre of the front side of the cube. For the cube without the Gabor pattern, noise patches are drawn in the same area as the Gabor patch and consisted of uniformly random noise pixels at 10% contrast, modulated by a Gaussian envelope. Right after the second cube disappeared a white question mark appeared and stayed on the screen until the participant made a key response. Accuracy was stressed but the participant was asked to give a key response in 3 seconds. The Matlab Quest Toolbox was used to adjust the contrast of the Gabor stimuli for each trial to identify a contrast value that leads to a performance level of 75% correct (Seow & Fleming, 2019; Watson & Pelli, 1983).

Belief task

After a short break, the participant performed the belief task. They were told that for each trial they were going to watch a short video, which involves an actor, a table, a white occluder and a grey cube with Gabor pattern on its front side. Each video began with the actor facing towards the table and the cube on the table in front of the occluder. After 1s, the cube started to move to a location behind the occluder. By manipulating the order of the events in the video (cube motion, actor motion), four belief conditions were created in a within-participant design (see Figure 2-1 and stimuli section). At the end of each video, right after the white occluder disappears and while the actor remains watching the table, a test cube appears in the middle on the table for 500ms. The test cube has the same colour (128, 128, 128) and size as the moving cube. On half trials, the test cube has a Gabor pattern on the front side and on the other half of the trials, the test cube has a white noise pattern on its front side. Unlike previous studies (Kovács et al., 2010), in the current study there was always a cube present at the end of the trial when the occluder was removed. This was because trials without a test cube cannot be analysed in our SDT approach, and so including these trials would add excess time to the study without adding useful data.

The contrast of the Gabor pattern on the test cube was determined by the previous thresholding test phase, and was selected so the participant should be able to achieve a 75% correct rate of response. The parameters of the noise dots were the same as the threshold testing stage. When the cube disappeared, a question mark showed up and stayed on the screen until response. Accuracy was stressed to the participants but they needed to respond in 3s. To enable SDT analysis, it is important that participants provide approximately 50% ‘yes’ and 50% ‘no’ responses in judging if the Gabor pattern is present on the test cube. To encourage this, participants were told explicitly at the beginning of the belief task that the proportions of Gabor pattern and noise patches were each 50/50. More importantly, if the script detected there was a response bias within a block (i.e. the number of YES responses is more than 16 or less than 8 in a block with 24 belief trials), the script would remind the participant during the next break to make approximately even numbers of yes/no judgements. Participants who persistently gave biased responses were excluded from the final analysis.

Overall, there were 4 belief conditions, each with 2 outcomes (either ending with a Gabor cube or with a noise cube), resulting in 8 types of trial. Each type of trial was repeated 24 times to allow enough trials for SDT analysis.

Attention check trials

To make sure that the participants were paying attention to the videos all the time rather than only to the last frame, we inserted another 48 trials as attention check trials. In the attention check trials, participants viewed shorter videos edited from the 4 belief conditions, and then answered a question either about the location of the cube or the orientation of the actor. There were 5 distinct versions of the attention check trials, each appearing randomly within each block. The accuracy of these questions was stressed to each participant and a low accuracy on these trials indicates participants were not paying enough attention to the videos so their data were excluded from further analysis.

In total, there were 240 trials for the whole belief task and 30 trials per block. Within each block, there were 6 trials from each belief condition (3 trials with a Gabor cube and 3 trials with a white noise cube) and 6 attention check trials. Belief trials and attention check trials were mixed randomly within each block. This manipulation helped participants to monitor their key responses and hence guarantee approximately equal numbers of 'yes/no' responses. Between blocks, there was a rest interval of at least 30s. The belief task took about 85min.

Post-experiment questionnaire

After completing all the computer-based tasks, the participants completed two questionnaires. The first one was related to the current task they performed, and mainly concerned their subjective report on their attention and knowledge about the actor, e.g. 'I paid a lot of attention to the actor,' and 'I wonder why she turns a lot.' Other questions are generally related to the participants' evaluation of the quality of the movie and the estimation of the task performance. The other questionnaire was the AQ (Baron-Cohen et al., 2001). The whole experiment took about 2hrs to be finished for each participant.

Data analysis

As specified in our pre-registration, we took the Gabor pattern detection data for the 40 valid participants and calculated *d'* and *criteria* values (Stanislaw & Todorov, 1999). The SDT *d'* was calculated by subtracting the z score for false alarm rate from the z score for hit rate (see formula 1) and criterion was calculated using the formula (2) listed below (Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999). These values were then submitted to a repeated measures ANOVA to test if the beliefs of the actor can bias the judgements of the participant, and thus could provide evidence for altercentric intrusion in this task.

$$d' = Z_{\text{hit}} - Z_{\text{false alarm}} \qquad (1)$$

$$c = -\,0.5 * (\, Z_{\text{hit}} + Z_{\text{false alarm}}) \qquad (2)$$

When preparing for this study, we did not anticipate having to exclude as many as 26 participants from our final data set[1], in 11 of these the exclusion was only because they gave unequal responses (e.g. all NO or all YES in a block) which cannot be used in a traditional SDT analysis. Therefore, we also decided to conduct exploratory analyses which could make use of more of the data which we collected, omitting items (3) and (4) in our original exclusion criteria and including all 55 participants who passed the attention check and scored within the typical range on the AQ. First, we analyzed reaction time data to test if there were differences related to either the participant's or the actor's belief, even though the task was not speeded. Here, we excluded trials where participants responded in less than 150ms, and those where participant's RT is below or above three standard deviations were also excluded from further analysis. Then the mean RT for each condition and each participant was submitted to a repeated-measures ANOVA.

Second, we adopted a Bayesian approach which is less influenced by extreme *d'* and *criteria* values. The BayesSDT model provides us a number of advantages:1) avoid edge corrections applied in standard SDT analyses that lead to biases in *d'* and *c* estimation when cell counts contain zeros; 2) allow group-level estimates of *d'* and *c* to mutually constrain extreme single-subject parameter estimates (Lee, 2008; Kruschke, 2010; Pleskac, Cesario, & Johnson, 2018). By performing this analysis we are able to obtain the posterior distributions of regression coefficients encoding the influence of our 2x2 factorial design on group-level

[1] 11 participants failed on attention check criterion (item 2); 14 participants failed because of low overall accuracy (item 3); 11 participants failed because of giving too many biased responses (item 4). Some of the participants failed on more than one criteria.

sensitivity and criterion parameters. The SDT model was nested inside a regression model that encoded the two factors of our experimental design (participant beliefs × actor beliefs), plus their interaction. Thus each subject's *d'* parameter was specified as:

$$d' = d'_{\text{base}} + \beta_{\text{pb}} * I_{\text{pb}} + \beta_{\text{ab}} * I_{\text{ab}} + \beta_{\text{i}} * I_{\text{pb}} * I_{\text{ab}}$$

where I_{pb} and I_{ab} are indicator variables that are equal to 1 when participant/actor is holding a target present belief and -1 otherwise, and β_{pb} , β_{ab} and β_{I} are regression coefficients encoding the effects of participants beliefs, actor beliefs and their interaction, respectively. Uninformative (high variance) priors on these influences on *d'* were specified as follows (after JAGS convention, variances are written as precisions or the reciprocal of variance):

$$d'_{\text{base}} \sim N\,(0, 0.001)$$
$$\beta_{\text{pb}} \sim (0, 0.001)$$
$$\beta_{\text{ab}} \sim (0, 0.001)$$
$$\beta_{\text{i}} \sim (0, 0.001)$$

Analogous models and parameter estimation were also applied for the criterion, *c*. Markov Chain Monte Carlo (MCMC) implemented in JAGS in R was used to draw samples from the posterior distributions. When calling JAGS, we implemented 2000 adaptation steps, 5000 burn-in samples and 50000 effective samples. For each parameter we run 3 chains and convergence of all chains was assessed both visually and using Gelman & Rubin's potential scale-reduction statistic *R* for all parameters (Gelman, & Rubin, 1992). Our average *R* was 1.00 with a maximum value of 1.01, indicating good convergence.

The posterior distributions of each parameter returned by JAGS were then directly used for Bayesian inference. For each indicator variable, we calculated the probability that the coefficient is smaller than zero. For example, P (β_{PB}) for *d'* stands for the probability that regression coefficients for participant belief on *d'* is smaller than zero. To distinguish these probabilities from classical P-values, we denote them as P_{θ}.

By implementing Bayesian hierarchical analyses we are therefore able to use data from all 55 participants who pass our basic checks to evaluate whether the agent's belief impacts on participant's performance. In the 'Results' section, we clearly signpost our analysis as 'pre-registered' or 'exploratory' to facilitate understanding.

2.4 Results

Pre-registered analysis

Following the pre-registered document, the analyses below are based on data from 40 participants who met all our inclusion criteria.

d' Analysis

One critical prediction in this study is that an agent's belief may influence the participant's perceptual sensitivity. To test this hypothesis, *d'* for each participant from each belief condition was calculated by using formula (1) listed above (Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999), then we submitted *d'* from all four conditions to a repeated-measures analysis of variance (ANOVA) with Agent's belief (target present, target absent) and Participant's belief (target present, target absent) as within-subject factors. Before statistical analyses, the sphericity of the data was verified (Mauchly's test, all p > .05).

Results showed no main effect for agent's belief (F (1, 39) = 0.034, p = 0.855) and participant's belief (F (1, 39) = 0.135, p = 0.715). Also no interaction was found between these two factors (F (1, 39) = 0.592, p = 0.446) (Figure 2-3.A).

Criteria Analysis

The SDT criterion was calculated using the formula (2) listed above (Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999). Criteria for each participant from each condition were also submitted to a repeated-measurement ANOVA with Agent's belief (target present; target absent) and Participant's belief (target present, target absent) as within-participant factors. Interestingly, results revealed a trending main effect for Agent's belief (F (1, 39) = 3.985, p = 0.053, η_{p^2} = 0.093).

Participant's criteria became less positive when the agent believed the cube was behind the occluder, indicating there is a weak trend to be more likely to report target presence in this condition. There was no main effect for participant's belief (F (1, 39) = 1.725, p = 0.197) and no interaction between Agent's belief and participant's belief (F (1, 39) = 0.109, p = 0.743) (Figure 2-3.B).

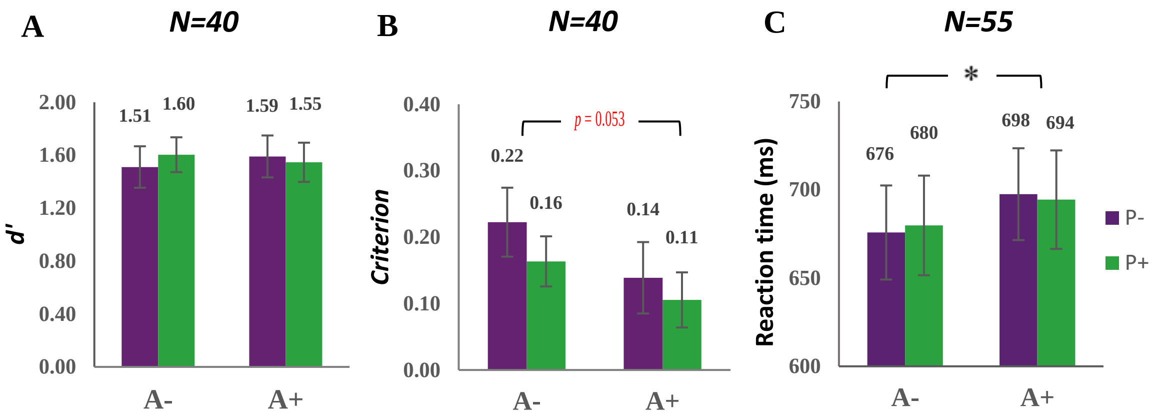


Figure 2-3. Mean d', criterion and reaction time between four belief conditions.

<u>*Exploratory analysis*</u>

These analyses include all 55 participants who completed the task and passed the attention check, and does not exclude those with biased responses.

RT analysis

RT data from all four conditions were submitted to a repeated-measures analysis (ANOVA). Results revealed a significant main effect for Agent's belief (*F* (1, 54) = 4.971, *p* = 0.030, η_{p^2}= 0.084). Participants responded slower when the agent believed there was a target cube. No main effect was found for participant's belief (*F* (1, 54) = 0.003, *p* = 0.959) or for the interaction between Agent's belief and participant's belief (*F* (1, 54) = 0.319, *p* = 0.575) (Figure 2-3.C). It is worth noting that similar results were also found when analysing data from the 40 participants who met all criteria we set out in the pre-registration.

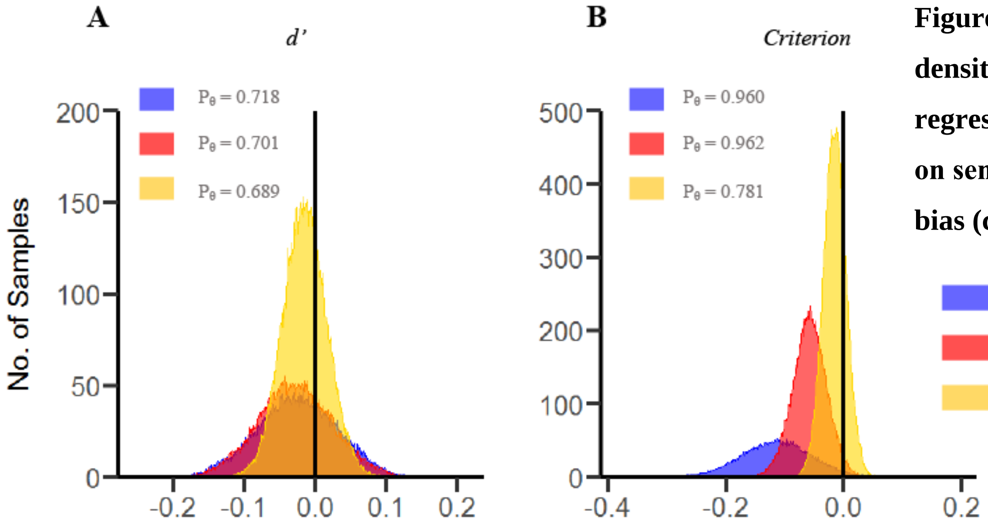


Figure 2-4. Posterior densities of estimated regression coefficients on sensitivity (d') and bias (criterion).

Bayesian SDT Analysis

The results from the Bayesian analysis engendered larger probabilities for impact both from participant beliefs and actor beliefs on the criteria but not on d'. Taking the d' analysis first (Figure 2-4.A). Bayesian analysis provides weaker support for positive influence (positive coefficients) either from participant beliefs (P_θ =0.718) or actor beliefs (P_θ =0.701) on performance with this measure, consistent with our classical pre-registered analysis. However, as Figure 2-4 shows, both factors are much likely to impact on criteria. When participants believe the cube is present, criteria are highly probable to decrease compare with when participants believe the cube is absent (P_θ =0.960) and similar trend is also observed when actor believes the cube is present (P_θ =0.962). But the probability for the interaction between these two factors is relatively smaller (P_θ =0.781).

2.5 Discussion

In this study, we designed a feature-detection false belief task to test whether typical adults spontaneously attribute beliefs to a mere co-observer. Signal detection theory (SDT) was applied to test if there is altercentric intrusion from another person's belief influencing a participant's perceptual process, with perceptual discrimination (*d'*) and a decision criterion (c) symbolizing different perceptual components. Our pre-registered analysis (n=40) hinted at a small effect of the actor's belief on the participant's decision criteria but this was not significant. Our exploratory Bayesian analysis revealed that both the participant's belief and the agent's belief can change the decision criteria. When either participants or the actor believed the target was present, participants turn to give more liberal and slower responses. We discuss these results first in terms of the interpretation of the decision criterion value, and then consider the robustness of implicit theory of mind effects.

2.5.1 Measuring implicit ToM with signal detection

In a signal detection framework, two measures of perceptual discrimination performance can be obtained. *d'* gives a measure of a participant's perceptual sensitivity to signal occurrence in signal-to-noise units, and the *criterion* reflects an internal cut-off line above which a participant will regard the internal evidence to be strong enough to represent a signal (Stanislaw & Todorov, 1999; Wyart, Nobre, & Summerfield, 2012). Changes to both *d'* and criteria can be induced by cues preceding the stimulus. Perceptual sensitivity can be increased by location cues which drive participants to focus attention on a specific location or a relevance cue which reduces uncertainty about the upcoming stimulus. Changes in criteria can also be induced by location or relevance cues, but also by strategic factors such as increasing the reward available for one stimulus interpretation or by creating prior expectations (Downing, 1988; Fleming, Whiteley, Hulme, Sahani, & Dolan, 2010; Summerfield & de Lange, 2014; Summerfield & Egner, 2009; Summerfield & Koechlin, 2010; Whiteley & Sahani, 2008; Wyart et al., 2012). Criterion shifts are thought to reflect changes either at a perceptual or decisional stage of processing (Witt, Taylor, Sugovic & Wixted, 2015).

These differences between *d'* and *criteria* and how they can be manipulated may help us to understand the results we find in the present study, in relation to previous work using SDT in a social task. Seow & Fleming used SDT in the context of a visual perspective-taking task and found that both participant's *d'* and criterion changed as a function of the avatar's visual perspective. Specifically, participants were more sensitive and more liberal when the avatar could also see the target object. This altercentric intrusion occurs despite the fact participants were not asked about the avatar's point of view on those trials. In contrast, our study shows that participants may change their criterion under the influence of another person's belief, in the absence of changes in *d'*. That is, participants may use a more liberal threshold to decide if the Gabor pattern is present but they do not improve in sensitivity.

The discrepancy between results of the two studies suggests that visual perspective and belief may have different mechanisms in influencing an individual's decision making. In the VPT tasks, attention may be a critical factor affecting performance (Catmur, Santiesteban, Conway, Heyes, & Bird, 2016), and many SDT studies have revealed an enhancement in attention can boost signal detection sensitivity (Wyart, Nobre & Summerfield, 2012; Downing, 1988; Summerfield & Egner, 2009). In contrast, spatial attention is less relevant in the current Gabor detection task because the pattern always appeared at the same spatial location. This may explain why *d'* did not change in the current experiment.

Both our study and that of Seow et al. find evidence that altercentric intrusion may change the criterion people use to make their decision, with participants using a more liberal criterion when another person can see the same stimulus (Seow & Fleming) or when the other believes the Gabor is present (this study). One possible reason is that processing another's mental states changes our expectations about what signals are present. A number of SDT studies have shown that expectations and rewards can bias perceptual decision-making on both behavioural and neural levels (Summerfield & Koechlin, 2010; Fleming, Whiteley, Hulme, Sahani & Dolan, 2010; Wyarta, Nobrea & Summerfield, 2012; Summerfield & de Lange, 2014), that is, when participants expect a stimulus to be present (e.g. it was frequently present on previous trials), then they have a more liberal criterion to judge that it is present in the future. Such expectations may be built up from previous experience in the current task (e.g. seeing the cube move behind the barrier leads to an expectation that it will be there later, a basic object permanency effect). This is reflected in the finding that, in our exploratory Bayesian analysis, participants are more likely to judge the pattern is present when they themselves believe the cube is present. This occurs despite the explicit instruction that the prior probability of signal/noise at the test phase 50/50.

Our data also gives hints of an altercentric intrusion effect in affecting criterion shifts. That is, the other's belief (or maybe expectation) that the cube is present also leads participants to be more liberal in their judgements and to report that the pattern was present. In this context, the actor's expectation that the cube is present seems to bias participant's judgements. This is a spontaneous altercentric intrusion because participants were never asked to make any judgements about the other person's beliefs nor to consider what the other knew during the task.

2.5.2 Reaction time measures

We also recorded reaction time data in our task, even though participants were not instructed to respond fast. There was some evidence that considering the other's beliefs may cause an increase in processing time. Reaction times on trials where the actor had a 'target present’ belief were significantly slower compared to when the actor had a target absent belief, despite this being an unspeeded task. As we strictly controlled the timing of key events (when the cube was last seen by participant, when the actor turns towards the table), a plausible account to explain the increase in RT is that participants spontaneously processed actor’s belief. When the actor is holding a target present belief, the onset of the final testing cube may trigger participants to compare the current cube with the actor’s belief, which might delay the participant’s response.

2.5.3 Effects of the participant’s own belief

In our canonical analysis on 40 participants, there was no effect of the participant’s own belief on *d’*, criteria or on RT. However, our Bayesian analysis with a larger sample size (N=55) did reveal a highly probable influence on criteria from participant’s belief (P = 0.960). From the distribution of Beta coefficients for participant’s belief, we can see the influence from participant’s belief on criteria has a large variation across all participants, and this variation might explain why from canonical analyses we could not see a significant influence from participant’s own beliefs.

Surprisingly, we did not find any influence from the participant’s own belief on reaction times, despite this belief affecting the decision criterion. Such a result may be caused by our task design, as in previous studies the target object can be present or absent at the end of the trial, but in our task the object (cube) was always present in the end of each trial. Only the presence or absence of the Gabor pattern on the cube was varied. We kept the cube present on every trial because we could not collect any SDT judgements on cube-absent trials, and including such trials would make the experiment too long. However, the fact that the cube was always present meant that participants would not be surprised by the physical presence of the cube, and this may have reduced the self-belief effects on reaction time.

It is useful to consider here how our task relates to the idea of inherent processing limits in the capacity for implicit ToM. In particular, Low & Watts (2013) proposed that the iToM system can track the location of an object (is the cube behind the screen or not?) but cannot track the identity of the object (is this the Gabor cube or plain cube?). Framing the task in these terms makes it even more surprising that participants showed any evidence of taking the agent’s belief into account, because it suggests that the participants go beyond the basic limitations of the iToM system. However, if we reframe the task in terms of ‘tracking the location of the Gabor patch’ and consider all other objects as distractors, then it might be possible to explain performance within a more limited location tracking iToM system. Further work will be needed to distinguish these possibilities.

Overall, our data provide hints that participant's perceptual judgements and their reaction times can be influenced by the beliefs of an actor, even when the actor is irrelevant to the task. This is manifested particularly in a change in criterion, which may be related to an unfolding expectation of events in the trial. However, these effects were weak and often marginal. We consider this issue next.

2.5.4 Robustness in the study of implicit theory of mind

Many previous studies of implicit forms of theory of mind in adults have given rather ambiguous results, with some papers showing evidence that adults spontaneously consider the mental states of others (Schneider, Bayliss, Becker, & Dux, 2012; Nijhof, Brass, Bardi, & Wiersema, 2016; Dumontheil, Apperly, & Blakemore, 2010; Schneider, Slaughter, Becker, & Dux, 2014; Van Der Wel, Sebanz, & Knoblich, 2014; Schneider, Nott, & Dux, 2014), while other papers argue against a pure iToM account (Catmur et al., 2016; Conway, Lee, Ojaghi, Catmur, & Bird, 2017; Phillips et al., 2015). Our results are also ambiguous, because our pre-registered analysis does not show an effect of the agent's belief (at p = 0.053 which is often considered marginal) and our exploratory Bayesian analysis does show an effect of agent's belief. Previous research suggests there are two plausible accounts for such a weak effect. One explanation is related with to which level ToM is measured in different tasks. As described above, it may be harder for an implicit mentalising system to track object identity, compared to object location (Low & Watts, 2013). As our task could be taken as an object identity task, this might explain why the results are weak. Another possible explanation is related with participants' social motivation to be involved in others mental states. Cane (2017) and Elekes (2016) both found in VPT research that participant's performance can be influenced by social factors such as monetary rewards, task instructions or the partner's task (Cane, Ferguson & Apperly, 2017; Elekes, Varga & Király, 2016). It's also worth noting that both of them found such effect with level-2 VPT tasks, which is quite similar to our study where object identity is tested. Following this vein, if we adopt more socially-engaged scenarios, participants may invest more effort to the task, so it's possible that we will observe a larger difference.

Given this pattern, we do not wish to draw definitive conclusions here about whether people do or do not engage in implicit theory of mind. Rather, we believe that our data suggests that it is worth continuing to study this area and looking for factors which may influence this process. That is, we should not abandon the domain of adult implicit theory of mind as a dead-end. Rather, our study builds on the work of Seow & Fleming (2019) in highlighting new ways to measure altercentric intrusions in adults. We believe that future studies using these methods may be able to provide definitive evidence in favour of the implicit mentalising hypothesis, and that such studies are worth pursuing. In particular, the signal-detection approach may provide a more sensitive measure than reaction times, and could be employed in a wider range of contexts in future.

Second, we suggest that it may be important to consider a wider range of factors when designing stimuli for studies of implicit mentalising. That is, it is not clear if implicit mentalising can be induced by any humanoid stimulus which moves in a human-like way, or if factors such as the animacy of the agent and the motivation of the observer to engage can change whether or not implicit mentalising is engaged. It may be that implicit mentalizing is more likely to occur when a richer social context is built as participants will be more motivated to be engaged in these tasks, but previous studies have not examined this. Both our study and Seow's study suggest that the application of SDT can help us decompose the perceptual process, and it will be worth exploring with SDT to ask how exactly these social factors influence the interaction between others' and our own mental contents.

2.6 Conclusion

In this study, we incorporated signal detection theory into a false belief task intending to explore in a minimal social context whether a co-observer's belief could influence our perception. We reveal that perceptual sensitivity was not influenced by other's belief contents; however, we find that the decision-making process was influenced by whether another person believes the target it present or not. When a co-observer believes a target is present, our decisions are slower and more liberal, but this influence from another's belief represents a weak effect. Compared with previous paradigms used in investigating implicit ToM, the method of the current study is more directly related to the belief representations in mind and the results provide more details on how other beliefs influences perceptual judgements. Future studies could introduce more social cues to seek to create stronger contextual effects of others' beliefs.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states?. Psychological review, 116(4), 953. https://doi.org/10.1037/a0016923

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844. https://doi.org/10.1111/j.1467-9280.2006.01791.x

Back, E., & Apperly, I. A. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition*, *115*(1), 54–70. https://doi.org/10.1016/j.cognition.2009.11.008

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, *14*(3), 110-118. https://doi.org/10.1016/j.tics.2009.12.006

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism Spectrum Quotient : Evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Cane, J. E., Ferguson, H. J., & Apperly, I. A. (2017). Using perspective to resolve reference: The impact of cognitive load and motivation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 591. https://doi.org/10.1037/xlm0000345

Carruthers, P. (2015). Mindreading in adults: Evaluating two-systems views. *Synthese*, 1–16. https://doi.org/10.1007/s11229-015-0792-3

Catmur, C., Santiesteban, I., Conway, J. R., Heyes, C., & Bird, G. (2016). Avatars and arrows in the brain. *NeuroImage*, *132*, 8–10. https://doi.org/10.1016/j.neuroimage.2016.02.021

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive development*, *9*(4), 377-395. https://doi.org/10.1016/0885-2014(94)90012-4

Conway, J. R., Lee, D., Ojaghi, M., Catmur, C., & Bird, G. (2017). Submentalizing or mentalizing in a Level 1 perspective-taking task: A cloak and goggles test. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 454–465. https://doi.org/10.1037/xhp0000319

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach*.

John Wiley & Sons.

Downing, C. J. (1988). Expectancy and Visual-Spatial Attention: Effects on Perceptual Quality. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(2), 188–202. https://doi.org/10.1037/0096-1523.14.2.188

Dumontheil, I., Apperly, I. A., & Blakemore, S. J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, *13*(2), 331–338. https://doi.org/10.1111/j.1467-7687.2009.00888.x

Elekes, F., Varga, M., & Király, I. (2016). Evidence for spontaneous level-2 perspective taking in adults. *Consciousness and Cognition*, *41*, 93-103. https://doi.org/10.1016/j.concog.2016.02.010

Fleming, S. M., Whiteley, L., Hulme, O. J., Sahani, M., & Dolan, R. J. (2010). Effects of Category-Specific Costs on Neural Systems for Perceptual Decision-Making. *Journal of Neurophysiology*, *103*(6), 3238–3247. https://doi.org/10.1152/jn.01084.2009

Frith, C. D. (1999). Interacting Minds--A Biological Basis. *Science*, *286*(5445), 1692–1695. https://doi.org/10.1126/science.286.5445.1692

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology*, *15*(17), R644–R645. https://doi.org/10.1016/j.cub.2005.08.041

Furlanetto T, Becchio C, Samson D, Apperly I. (2015). Altercentric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. *Geologia Tecnica e Ambientale*, *19*(3), 55–79. https://doi.org/10.1016/j.jfms.2010.11.013.

Gelman, A., Rubin, D. B., Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences Linked references are available on JSTOR for this article : Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. http://doi.org/ 10.1214/ss/1177011136

Herman, L. M., & Kantowitz, B. H. (1970). The psychological refractory period effect: Only half the double-stimulation story? *Psychological Bulletin*, (May). https://doi.org/10.1037/h0028357

Heyes, C. (2014). Submentalizing. *Perspectives on Psychological Science*, *9*(2), 131–143.

https://doi.org/10.1177/1745691613518076

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*, *79*(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kovács, A. M., Teglas, E., & Endress, A. D. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, *330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Kovács, Á. M., Kühn, S., Gergely, G., Csibra, G., & Brass, M. (2014). Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PloS one*, *9*(9). https://doi.org/10.1371/journal.pone.0106558

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300. https://doi.org/10.1016/j.tics.2010.05.001

Lee, M. D. (2008). BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods*, *40*(2), 450–456. https://doi.org/10.3758/BRM.40.2.450

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, *8*(12), 528–533. https://doi.org/10.1016/j.tics.2004.10.001

Low, J., & Perner, J. (2012). Implicit and explicit theory of MD: State of the art. *British Journal of Developmental Psychology*, *30*(1), 1–13. https://doi.org/10.1111/j.2044-835X.2011.02074.x

Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science*, *24*(3), 305-311. https://doi.org/10.1177/0956797612451469

Naughtin, C. K., Horne, K., Schneider, D., Venini, D., York, A., & Dux, P. E. (2017). Do implicit and explicit belief processing share neural substrates?. *Human brain mapping*, *38*(9), 4760-4772. https://doi.org/10.1002/hbm.23700

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin*, *89*(1), 133–162. https://doi.org/10.1037/0033-2909.89.1.133

Nijhof, A. D., Brass, M., Bardi, L., & Wiersema, J. R. (2016a). Measuring mentalizing ability: A within-subject comparison between an explicit and implicit version of a ball detection task. *PLoS ONE*, *11*(10), 1–15. https://doi.org/10.1371/journal.pone.0164373

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?. *Science, 308(5719)*, 255-258. http://doi.org/ 10.1126/science.1107621

Phillips, J., Ong, D. C., Surtees, a. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A Second Look at Automatic Theory of Mind: Reconsidering Kovács, Teglas, and Endress (2010). *Psychological Science*, 1–15. https://doi.org/10.1177/0956797614558717

Pleskac, T. J., Cesario, J., & Johnson, D. J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin and Review*, *25*(4), 1301–1330. https://doi.org/10.3758/s13423-017-1369-6

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review, 12(2)*, 195-223. https://doi.org/10.3758/BF03257252

Samson, D. (2013). *Theory of Mind*. *Encyclopedia of Identity*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0059

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & 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. https://doi.org/10.1037/a0018729

Santiesteban, I., Catmur, C., Hopkins, S. C., & Bird, G. (2014). Avatars and arrows: Implicit mentalizing or domain-general processing? *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 929–937. https://doi.org/10.1037/a0035175

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*(3), 433–438. https://doi.org/10.1037/a0025458

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological science*, *23*(8), 842-847. https://doi.org/10.1177/0956797612439070

Schneider, D., Nott, Z. E., & Dux, P. E. (2014). Task instructions and implicit theory of mind. *Cognition*, *133*(1), 43–47. https://doi.org/10.1016/j.cognition.2014.05.016

Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage*, *101*, 268–275. https://doi.org/10.1016/j.neuroimage.2014.07.014

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science, 325*(5942), 883-885. http://doi.org/ 10.1126/science.1176170

Seow, T., & Fleming, S. M. (2019). Perceptual sensitivity is modulated by what others can see. *Attention, Perception, & Psychophysics, 81*(6), 1979-1990.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587-592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756. https://doi.org/10.1038/nrn3838

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409. https://doi.org/10.1016/j.tics.2009.06.003

Summerfield, C., & Koechlin, E. (2010). Economic Value Biases Uncertain Perceptual Choices in the Parietal and Prefrontal Cortices. *Frontiers in Human Neuroscience*, *4*(November), 1–12. https://doi.org/10.3389/fnhum.2010.00208

Van Der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition, 130*(1), 128–133. https://doi.org/10.1016/j.cognition.2013.10.004

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method.

Perception & Psychophysics, 33(2), 113–120. https://doi.org/10.3758/BF03202828

Whiteley, L., & Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision, 8*(3), 2. https://doi.org/10.1167/8.3.2

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128. https://doi.org/10.1016/0010-0277(83)90004-5

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception, 44(3)*, 289-300. https://doi.org/10.1068/p7908

Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, *109*(9), 3593–3598. https://doi.org/10.1073/pnas.1120118109

Chapter 3. Spontaneous theory-of-mind in an interactive social context

3.1 Introduction

In Chapter 2, we examined spontaneous theory-of-mind (ToM) in a minimal social context, i.e. where there is no interaction among the self-other(s)-object(s) triadic relationship. We combined the signal detection theory with a false belief task, and asked participants to detect near-threshold Gabor stimuli on a box after watching a video when participants and the actor in the video held true or false beliefs on whether the box is present or not. Results showed that in this minimally social environment, participants' decision criteria were subtly influenced by the actors' beliefs. When the actor believed the cube was present, participants had a stronger tendency to report that they saw the Gabor pattern, and such decisions were more time-consuming than that in other conditions.

However previous studies on spontaneous ToM usually report larger effects (Senju et al., 2010; Southgate, Senju, & Csibra, 2007; Schneider, Bayliss, Becker, & Dux, 2012; Kovács, Teglas, & Endress, 2010). A potential reason for the different results may lie in the distinct features in the experimental designs between our paradigm and the previous studies. As stated in Chapter 1, in almost all previous research on spontaneous ToM, the agent would have some simple interactions with the target object. Although such interactions might not be designed on purpose, they are significant to how participants process these social events as coherent scenes, and might change the extent to which they engage in mentalizing. On the contrary, in our minimal social settings in Chapter 2, the agent has no interaction with the object, thus it is possible that participants treated the agent and the target as disconnected parts of the scene which consequently reduces their propensity to engage in mentalizing.

The other-object interaction matters for social cognition more than simply attaching distinct items from a visual scene together. Similar to the Gestalt principles, for a long time philosophers and psychologists have argued that interaction allows for closer temporal or spatial relationships, or more items to be processed for our cognitive system (Baldassano, Beck, & Fei-Fei, 2017; David A. Rosenbaum, Kate M. Chapman, Matthias Weigelt & Wel, 2012; Kim, Biederman, & Juan, 2011; Klapp & Jagacinski, 2011). Instead, people can sense the coherence beyond the interactive components, and alter the way information is organized and stored in our mind. Kim et al. (2011) investigated the object-to-object interaction. They applied TMS and found that interactive pairs of objects were recognized better than the non-interactive pairs, but such boost effect disappeared when TMS applied to the lateral occipital cortex (LOC), indicating the benefit from interaction may relate with high-level visual processing. Ding and colleagues investigated interactive biological motions in working memory, and found that people can maintain the same amount of pairs of interactive biological movements (BMs) as individual BMs, suggesting human-to-human interaction can result in information compression thus saving cognitive resources (Ding, Gao, & Shen, 2017). Recently, Baldassano et al. (2017) directly investigated the neural mechanism of human-to-object interaction. They found that several brain regions such as the LOC, the extrastriate body area and the parahippocampal place area responded in a linear manner, adding up the agent and object in a scene, while the activation of the posterior superior temporal sulcus cannot be predicted by this simple addition. These results suggest that perceiving interaction can alter the way people make sense of the incoming information. It is likely that through observing the interactive components, people may represent individual items as a single unit, thus alter the strategy of maintaining or manipulating them in the cognitive system.

Unlike object-to-object interactions, human-to-object and human-to-human interactions are special because through observing the way other people interact with the object, and through interpreting their behaviours, one can acquire the critical intentional information of the agent. Such intentional information can enable an individual to communicate with others in a more effective and efficient way. In addition, interpreting human-to-human or human-to-object interaction is an important avenue for social learning. This goal is achieved both through attention and through imitation. Previous studies with infants showed that human-to-object interaction can draw their attention to the interactive components. In one study, researchers invited infants and their parents to the lab and recorded their interactive behaviour patterns during the experiment. They found that when infants saw their parents attending to the same object as they do, they attended to the object for a longer period and their retention of the object is better (Yu & Smith, 2016). In adults, researchers also found that interaction can facilitate attention-shift between the interactive components. In a new study, participants were presented with images of pairs of hands. Some pairs performed a handshake movement while others were random pairs. Results showed that participants' attention shifted more quickly between the two handshaking hands (Yin, Xu, Duan, & Shen, 2018). These novel results suggest that people are more involved when processing human-to-human or human-to-object interactions, indicating they have a high priority in the hierarchy of information.

Another way for people to gain social learning through watching interaction is imitation. When observing social interactions, individuals can spontaneously simulate the way other people interact with each other or with the object, and hence understand their social roles or identities, or the function of the objects. Previous research has shown that such mimicry behaviours are at least partially rooted in the mirror neuron system (MNS). Observing others manipulating an object activates the MNS and its adjacent areas (Catmur & Heyes, 2019; Errante & Fogassi, 2019), which has been suggested can play a significant role in movements rehearsal and recall. These previous findings supported the hypothesis that human-to-human or human-to-object interaction can constitute guidance for us to navigate the social world. Due to its significance, encountering social interactions may be one of the gateways through which we start mentalizing.

In the visual cognition domain, human-to-human or human-to object interactions are usually conveyed by goal-directed actions such as gazing, grasping and reaching. In previous research on mentalizing, almost all tasks included such cues of interaction. In the anticipatory-looking paradigm (Southgate et al., 2007; Senju et al., 2009; Schneider et al., 2012), participants observed the actor seeing or not seeing the change of location of the object. At the end of each experimental trial, or sometimes in the introductory videos, participants would see the actor reached the target, suggesting a desire for the object, which could trigger participants to predict the actor's behaviour when their eye-gaze was measured. Making such predictions needs an understanding of the agent's beliefs, thus participants were more likely to trace the actor's mental states and register their beliefs. Similarly in the Kovács' Smurf paradigm (Kovács et al., 2010), at the beginning of each experiment trial, the Smurf would place the target ball on the desk, and gaze toward it during its movement. Although the Smurf didn't reach the ball to get it in the end, the 'placing' and 'gazing' behaviours could suggest the participants that the agent cared about the location of the ball, which may increase participants' propensity to engage in mentalizing.

In this current study, we directly tested the hypothesis that human-to-object interactions may be one of the triggers for people to start considering other's mental states. In a context where a lot of agent-to-object interactions are present, participants would have more reasons to represent other's mental contents, so as to understand the agent's beliefs and predict their next move. To test our hypotheses, we modified the paradigm we used in Chapter 2, and designed comparable trials where the actor saw (or didn't see) the change of object's location in a more interactive or non-interactive context. To achieve higher ecological validity, we filmed all the experimental videos in a real scene and changed the target object to a more personal object—a programmable toy robot with the ability to move without external force. Specifically, in the interactive context, the actor would introduce the robot as a personal toy and display goal-directive movements such as pushing or reaching in the video to suggest that she was interacting with the robot. However in the non-interactive context, the agent would deny the robot was hers and behaved as a mere observer, similar to our manipulation in Chapter 2. The task for participants was to recognize the Gabor patterns on the toy robot. We hypothesize that in the non-interactive context, similar to the results in Chapter 2, the actor's belief may exert a weak influence on the participants' decisions. Thus, participants might adopt slightly more liberal criteria and respond faster when the actor believes the toy robot is present. But in a more interactive context, participants may engage more in ascribing the actor's beliefs, thus participants' decision criteria should be more liberal and reaction time should be shorter when the actor holds a target-present belief, compared with when the actor doesn't believe the robot is present.

3.2 Method

Based on our previous findings, we would like to investigate the hypothesis that in a more interactive social context, whether people would exhibit a stronger propensity to engage in other's mental state or not. Thus, we adopted a 2 (Context: interactive, non-interactive) × 2 (Participant's belief: presence, absence) × 2 (Agent's belief: presence, absence) within-subjects design, and created a set of new video stimuli to make the task more ecological validate. Specifically, the interactive condition and the non-interactive condition differ in two aspects: the background story at the beginning of each block and the motion cues presented in each trial. For the interactive condition, before each block, we would present participants with a narrative from the actor, in which she described explicitly her relationship with the toy robot, and expressed the follow-up videos were a display of an interactive game they used to play. Then in each trial, she would display some goal-directed actions such as push and reach to the toy. In the non-interactive condition, she expressed the toy was not hers and in the videos she would be a mere observer of the toy robot's movements.

Participants

55 right-handed adult participants took part in our experiment (28 females, 24.8 ± 5.7). Participants were required to have normal or corrected-to-normal vision, no hearing disability and have not participated in the experiment in Chapter 2 before. Participants were recruited from two UCL-based participant recruitment systems and were informed that the study would take place on two consecutive days. They were paid on the basis of £7.5 per hour to compensate for their time. This study has been approved by the UCL ethics research committee and follows the Declaration of Helsinki 1964.

Stimuli & Materials

Target object

To make sure the target object and move freely and smoothly in the video as the target cube in Chapter 2, we used a toy robot called Dash (https://uk.makewonder.com/dash/) and controlled its movements in different conditions by programming its movement trajectories before filming the videos. To present the Gabor stimuli, we made a special cube helmet with the same Gabor stripes presenting on the three sides of its head (left, right and back), and decorated Dash differently for the interactive and the non-interactive conditions. In the interactive condition, the bottom part of Dash was decorated with colourful stickers, and in the introductory video the participants were informed that the actor put the stickers on Dash and made it the striped helmet. For the non-interactive condition, the colourful stickers were replaced with white stickers. The Dash's movement trajectories were programmed with Blockly (Google & MIT) and the sound of its movement was removed when editing the videos.

Introductory videos

To make a direct comparison between the interactive and the non-interactive condition, we first filmed an introductory video for each condition to present explicit social information between the actor and Dash.

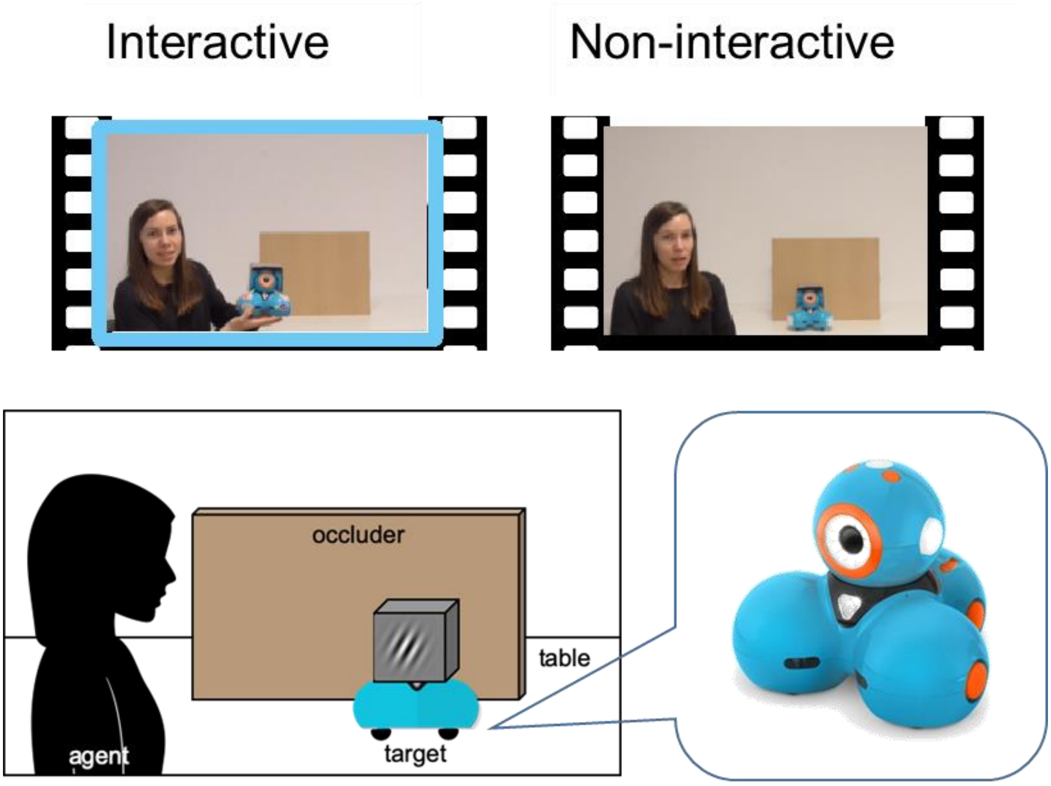
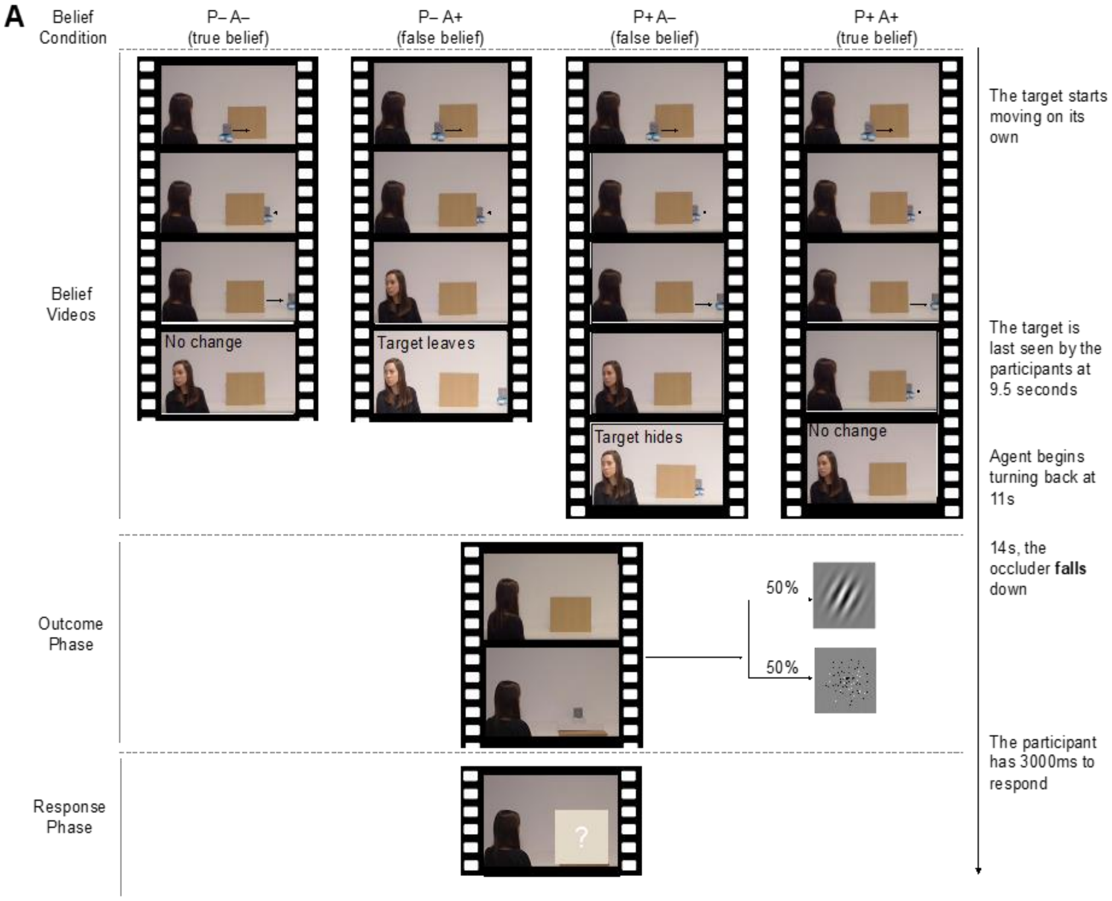


Figure 3-1. The demonstration of the experiment setting for filming the videos. The actor sat on the left with her body and head orienting towards a white table. The toy robot Dash always started moving from the left side of the table, and in the center of the scene a wooden occluder (about 50cm × 30cm × 5cm) was placed. The Dash's helmet was made from a 10cm × 10cm × 10cm cardboard box, wrapped with grey paper and on its three sides Gabor pattern were presented in a ~ 7cm × 7cm square. The Dash robot used in the experiment was shown in the picture on the right.

As part of the social context manipulation, short videos with sound, in which the actor speaks about the toy and whether she owns it, are used to establish the agent-object relationship, providing either a neutral or enriched social context. The actor presents the target as belonging to her in the interactive condition: she held Dash in her hands and showed the stickers on it, while saying: "Let me introduce you to my toy, Dash. I decorated it with stickers, and before we used to play hide and seek. Now let me show you how we play." In the non-interactive condition, the actor treated Dash as someone else's toy. For the whole video the toy is placed on the table, and the agent says: "This is not my toy. I don't know whose toy it is." (Figure 3-1).

Belief videos

Following Kovács' work and our previous design in Chapter 2, we created videos establishing the four belief combinations as shown below (P stands for participants, A for Agent, + for Dash present and – for Dash absent). Each belief video begins with the robot moving from the left of the table and hiding behind the occluder. Then we manipulated the orders of these events: 1) Dash moves out of the occluder 2) Dash moves back behind the occluder 3) The actor turns and faces away from the table 4) The actor turns back facing towards the table. Therefore we created different conditions where the actor or the participant's belief can either be Dash present (behind the occluder) or Dash absent, and the actor's belief can either be a true or false belief.


A
Belief Condition
P– A– (true belief)
P– A+ (false belief)
P+ A– (false belief)
P+ A+ (true belief)
Belief Videos
No change
Target leaves
Target hides
No change
The target starts moving on its own
The target is last seen by the participants at 9.5 seconds
Agent begins turning back at 11s
Outcome Phase
14s, the occluder **falls** down
50%
50%
Response Phase
?
The participant has 3000ms to respond

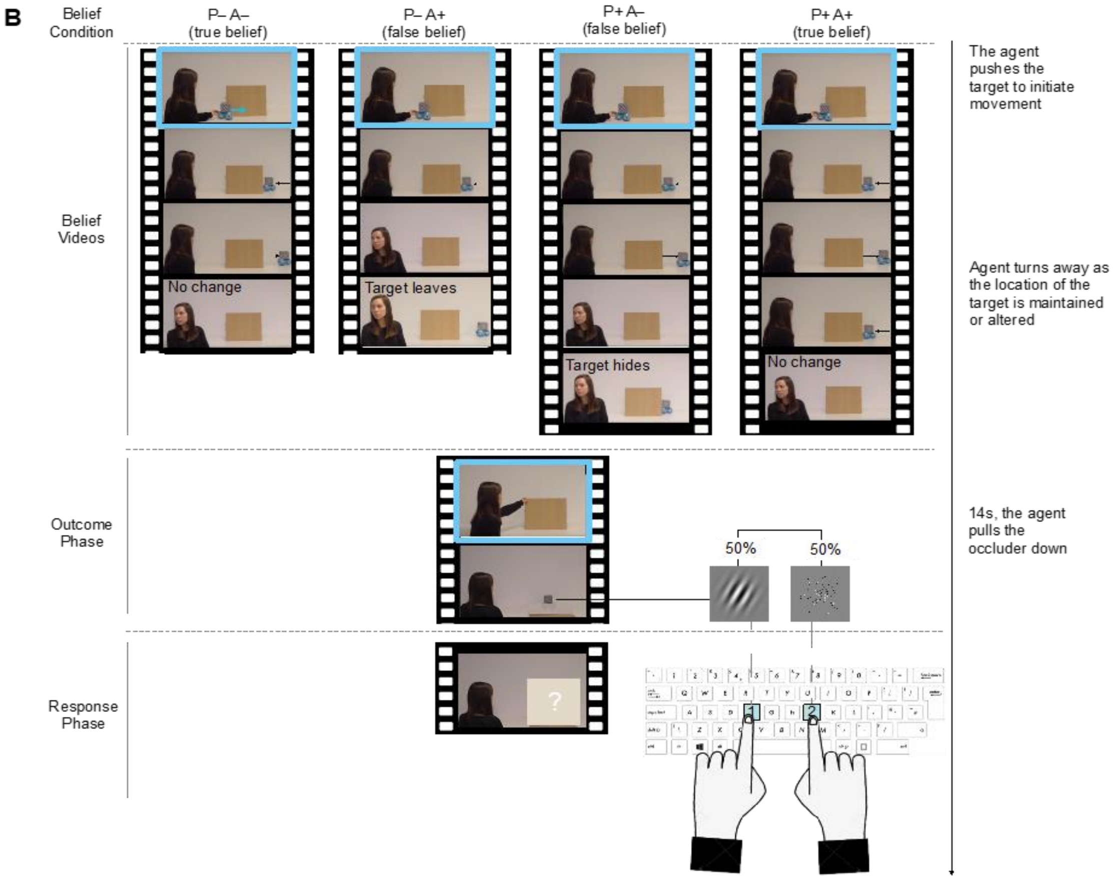


Figure 3-2. Frames from the introductory videos of the non-interactive (A) and interactive (B) conditions. In the introductory video of the non-interactive condition, the actor said this Dash was not hers as it didn't have the colourful stickers, and she didn't want to play with it. In the interactive condition, the actor held Dash in her hands and introduced Dash as her personal toy in her childhood. She also stated that she decorated Dash with her colourful stickers. At the end of the introductory video, she mentioned that in the following videos she was going to show a game she used to play with the robot.

1. P–A–: Dash moves behind the occluder and then moves out, all under the actor's gaze. Dash is last seen by the participant at 9.5 s. Then the actor turns away from the table and starts to turn towards the table at 11 s. The occluder falls down at 14 s. In this condition, both the participant and actor hold a "target-absent" belief;
2. P–A+: Dash moves behind the occluder. Then the actor turns away from the table. While she is not looking, Dash moves out of the occluder and out of the screen. Dash is last seen by the participant at 9.5 s. Then the actor turns towards the table at 11 s. The occluder falls down at 14 s. In this condition, the participant believes the target is absent but the actor has a false "target present" belief;
3. P+A–: Dash moves behind the occluder and then moves out of the occluder and out of the scene. Then it moves back behind the occluder again. The actor then turns away

from the table. Dash is last seen by the participant at 9.5 s. Then the actor turns towards the table at 11 s. The occluder falls down at 14 s. In this condition, the participant believes the target is present but the actor has a false "target-absent" belief;

4. P+A+: Dash moves behind the occluder and then moves out of the occluder and out of the scene. Then the actor then turns away from the table. While she is not looking, Dash moves back behind the occluder. Dash is last seen by the participant at 9.5 s. Then the actor turns towards the table at 11 s. The occluder falls down at 14 s. In this condition, both the participant and actor hold a "target present" belief;

Unlike videos in the non-interactive condition, in the interactive conditions, Dash starts moving as if it was pushed by the actor. In addition, at the end of each video, the actor would pull down the occluder showing she would like to see if Dash was there behind or not. However, in all videos under the two conditions, we aligned the time points of the key events (Dash's last seen by the participant, the actor's turning back and the occluder falls/pulled down), to exclude confounding factors such as movements or memory which is likely to influence reaction time (see Figure 3-2. for the timeline of videos in the interactive and non-interactive conditions). All the videos were edited using Adobe Premier Pro CC 2017 to align the timing of the key events and to adjust brightness and shadow (Phillips et al., 2015), then they are exported at 29 frames per second as .wmv files with 1024 x 576 pixel resolution.

The Last Frame (testing stage)

Closely following each belief manipulation video, we added a final frame in which participants needed to distinguish the Gabor or the noise pattern. For belief videos under each condition, the video frames immediately following the lowering of the occluder were extracted and saved as a static picture, and were appended to their parent videos seamlessly in Matlab 2018b. These static pictures constituted the background of the testing stage in each video, allowing us to superimpose the Gabor pattern or the white noise pattern onto the helmet of Dash. These last frame pictures were also used as the background when testing each participant's Gabor detecting threshold, to ensure the environmental consistency for Gabor feature detection across the experiment.

Attention check videos

To monitor participants' attention during the task without causing a confounding effect on our results, we inserted attention check trials as we do in Chapter 2. The attention check videos were edited from the belief manipulation videos, and were cut early before the occluder was lowered. Then in Matlab 2018b, we appended a blue screen in which a question related to several facts of the videos however irrelevant to belief was displayed. Questions predetermined were either about the current location of Dash (eg. Is Dash behind the occluder or out of the scene?), or the orientation of the actor (eg. Is the lady facing toward the left or right?), and two alternative choices were presented below the question to allow participants select their answers using the same keys to respond for the feature detection. For each Context condition, three attention check videos were created.

In the end, the final stimuli set included eight belief manipulation videos, eight corresponding final frames, and 6 attention check videos.

Procedure

The experiment took place over two consecutive days. This is because to conduct Bayesian Hierarchical Analysis, we need to guarantee enough trials (48 at least) under each condition, however, too many trials would make the experiment extremely long and including fatigue effect if each participant completed it in one day. Thus, we divided the experiment into two identical parts, each contained a Gabor threshold testing stage, then 6 blocks of the belief task, with half from the interactive condition and the other half from the non-interactive. Participants received the task information, signed the consent form, and finished the practice trials before each task on the first day, then on the second day after all tasks, they need to fill out the post-experiment questionnaires.

On both days, the participant completed the Gabor threshold testing then the belief tasks. All the stimuli are presented by Cogent 2000 and Cogent Graphics (http://www.vislab.ucl.ac.uk/cogent.php) toolboxes in MATLAB 2018b (The MathWorks, Natick, MA).

Threshold Testing

This phase of the study aims to titrate the Gabor detecting threshold for each participant for use in the second part of the study. Participants were asked to detect Gabor patterns of varying contrast in a 2 interval forced-choice task. The Gabor patterns were presented with contrasts varying from 0 to 1 on 56 × 56 pixel grey squares (148 × 148mm on-screen). As mentioned before, the test stimuli were then presented covering Dash’s helmet in a scene drawn from the last frame of the video clips, to ensure surrounding contrast and lighting were identical between the threshold testing and the belief task phases.

For each trial, participants were requested to detect which cube has the Gabor pattern, the first or the second. On the first day of testing, there were 10 practice trials before the formal threshold testing task, but we removed the practice trials on the second day as participants were familiar with the procedure of the task. Then participants were informed that the Gabor pattern would be difficult to detect during the formal session, such that they need to be concentrated. Participants were allowed to make the best guess when they felt unable to identify the Gabor pattern. There were 120 trials in the threshold testing session.

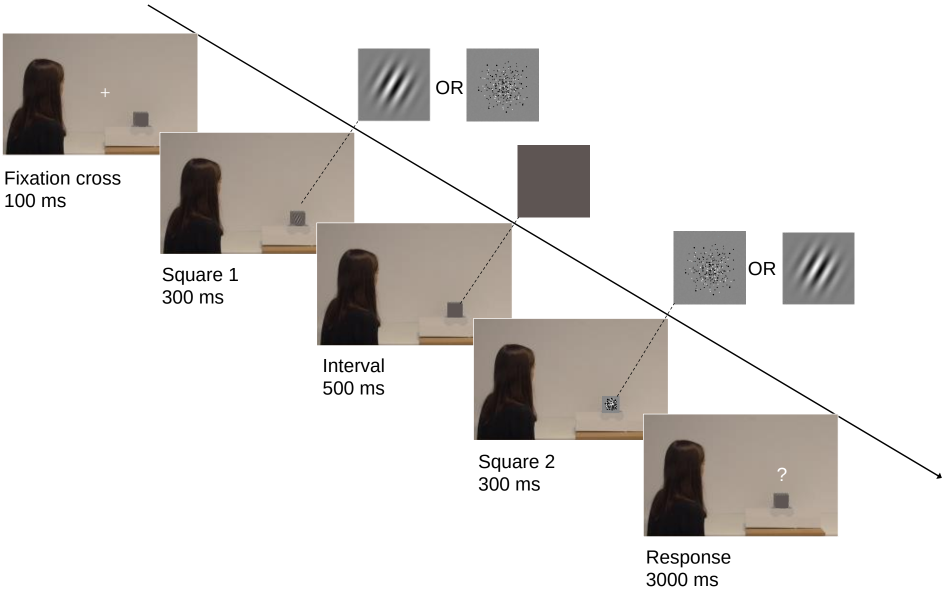


Figure 3-3. The timeline of a single trial from the threshold testing stage.

Each trial started with the onset of the last-frame picture (see *The Last Frame* in the *Stimuli & Material*) in which a female actor sitting in front of a table, with a wooden occluder lowered revealing the upper part of Dash with a grey helmet. The bottom part of Dash was masked only to reveal its silhouette. Then a white fixation appeared followed by the first grey cube, positioned over the helmet of Dash. On half of the trials, the Gabor pattern was presented on the first cube and for the other half on the second, subtending ~1.41 degrees of visual angle to the centre of the square. For the cube without the Gabor pattern, noise patches are drawn in the same area as the Gabor patch and consisted of uniformly random noise pixels at 10% contrast, modulated by a Gaussian envelope. Right after the second square disappeared, a white question mark appeared at the same location of the grey square, and the participant had 3 s to indicate whether the Gabor patch appeared first or second. Participants were instructed to press key ‘F’ if they thought the Gabor pattern was on the first grey square, and key ‘J’ if they thought it was on the second. Accuracy was stressed but the participant was also asked to respond as fast as possible (Seow & Fleming, 2019; Watson & Pelli, 1983) The Matlab Quest Toolbox was used to adjust the contrast of the Gabor stimuli for each trial to identify a contrast value that leads to a performance level of 75% correct (Seow & Fleming, 2019; Watson & Pelli, 1983) (Figure 3-3.).

Belief task

After the threshold testing stage, participants were introduced to the belief task. Participants were told that in this part their task was still to detect the Gabor pattern, but for each trial they would first watch a video clip then make the judgment. Verbal instructions were given with a picture demonstrating the setup of the videos. Participants were informed that each video would show a lady in a room with a robot moving around. At the end of each video, they needed to report if they saw the Gabor pattern on the helmet of the robot, by pressing ‘F’ if they saw the Gabor stripes and ‘J’ if they didn’t. Participants conducted 10 practice trials before the task to get familiar with the procedure.

To compare spontaneous ToM in the different social contexts, the whole belief task was structured into 12 blocks, with 6 blocks containing trials from the interactive condition (interactive blocks) and another 6 with trials from the non-interactive condition (non-interactive blocks). Each block began with an introductory video. For the interactive blocks, in the introductory the actor held Dash in her hands, stated that Dash is her toy and she decorated Dash with the colourful stickers. Then she said in the next videos she was going to display a game she used to play with Dash. Each trial in the interactive condition started with the actor sitting facing the table, and Dash stood next to her on the table. At the beginning of each video clip, she gave Dash a gentle push then Dash started moving to the right, hiding behind the occluder, then depending on each condition, Dash and the actor would do different movements. At the end of each video, the actor always turned back facing towards the table at 11s and reached out to lower the occluder, indicating she would like to see if Dash was behind it or not. All videos stopped after her hand touched the occluder at 14s.

For the non-interactive condition, Dash was decorated with white stickers to make it distinguishable from it in the interactive condition. In the introductory video, the actor expressed that the robot was not Dash with the robot placing on the table, and she didn't want to play with it. Then in each trial, she sat facing the table as a mere observer, without any interaction with the robot. Each video stopped at 14 s.

At the end of each video, the last frame picture extracted from each of its parent video was appended seamlessly to it, in which the actor remains watching the table, a grey square containing either the Gabor pattern or the white noise patch was superimposed onto Dash's helmet and was presented for 500 ms. The bottom part of Dash was masked only to show its silhouette so that the stickers on it wouldn't influence participants' judgments. The grey square applied onto Dash's helmet has the same colour (128, 128, 128) and size as the helmet. On half trials, it contained the Gabor pattern and on the other half with white noise patch. After 500 ms, an opaque beige square appeared to mask the whole figure of Dash and a question mark showed up which suggest that participant need to make a response. The question mark remained on the screen until a response was given or until 3 s. Using the same design in Chapter 1 but unlike previous studies (Kovács et al., 2010), Dash was always behind the occluder at the end of the trial when the occluder was lowered. This was because trials without a test cube cannot be analysed in our SDT approach, and so including these trials would add excess time to the study without adding useful data.

The contrast of the Gabor pattern on the grey square in the outcome phase was determined by the previous thresholding test phase, and was selected so the participant should be able to achieve a 75% correct rate of response. The parameters of the noise dots were the same as the threshold testing stage. To enable SDT analysis, it is important that participants provide approximately 50% 'yes' and 50% 'no' responses in judging if the Gabor pattern is present on the test cube. To encourage this, participants were told explicitly at the beginning of the belief task that the proportions of Gabor pattern and noise patches were each 50/50. More importantly, if the script detected there was a response bias within a block (i.e. the number of YES responses is more than 22 or less than 10 in a block with 32 belief trials, at .05 significant level), the script would remind the participant during the next break to make approximately even numbers of yes/no judgments. Participants who persistently gave biased responses were excluded from the final analysis.

To ask participants to maintain attention on the task, for each block we randomly inserted three attention check trials (see Stimuli & Materials). Four each interactive or non-interactive block, belief videos under each belief condition (P+A+, P+A-, P-A+, P-A-) were repeated for 8 times with half paired with Gabor pattern and the other half with noise patch in the end. In total, each block contains an introductory video, 32 belief trials balanced on conditions & outcomes, and 3 attention check trials. The belief trials and attention check trials were presented in a fully randomised order.

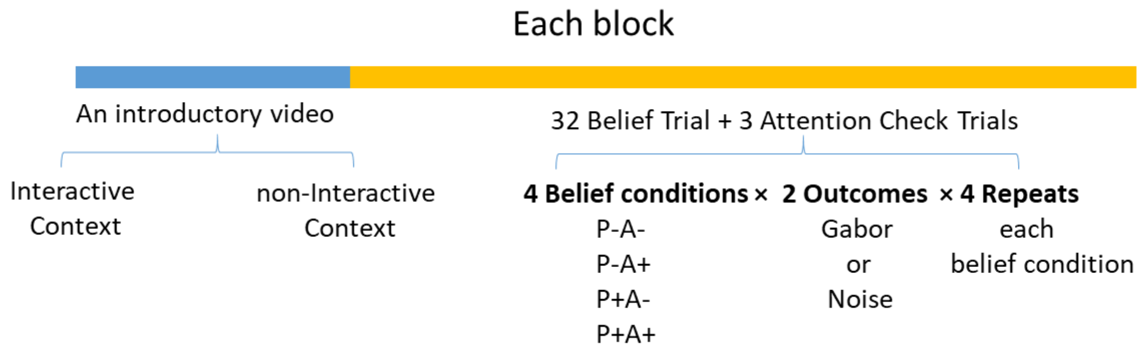


Figure 3-4. The trial organization structure of the task.

The belief task is composed of 48 trials under each Context (interactive, non-interactive) × Participant's belief (Dash presence, Dash absence) × Agent's belief (Dash presence, Dash absence) conditions, with another 32 attention check trials (Figure 3-4.).

Post- experiment Questionnaires

Participants were requested to fill out two questionnaires after completing all tasks. The first questionnaire checked their preference & attitude towards the actor, and their explicit knowledge about the actor's beliefs. The second questionnaire is the Autism-Spectrum Quotient (AQ) (Baron-cohen et al., 1985). The AQ consists of 50 questions, with each question having the value of 1 if it corresponds to an autistic trait and 0 if not, resulting in a maximum score of 50.

Between two blocks, there was a rest break for at least 30s. The whole experiment took about 2.5 hrs to finish with 1.25 hrs on each session.

Data Analysis

Datasets of the same participant from the two-days session were first merged for further analyses. Corresponding to data analysis methods in Chapter 2, we set these data exclusion criteria to ensure the robustness of the results.

The participant's data were excluded from the final analysis if:

1. The dataset was biased. Following the Chi-squared test, if participants gave more than 22 or less than 10 'signal' responses (out of 32 trials in one block), it is considered that participant's responses are not evenly distributed according to a 50/50 distribution (at a significant level of .05), thus this block was marked as a biased block. Participants with more than 3 biased blocks were excluded from the final analysis.
2. Participants scored above 32 on the AQ. Participants scored above the cut-off 32 points on AQ suggested that they are likely to have more than average autistic traits thus may influence their propensity in spontaneous ToM (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001a).
3. Overall accuracy on belief trials exceeded 0.95 or fell below 0.55.

Note that we set up the attention check trials for the task to monitor if participants paid enough attention to the task. However when analysing the data in the end, we found that some participants had extremely low accuracy in these trials. Looking back on the task design we found that questions related to the actor's orientation were very confusing to the participants as they could answer based on a different frame of reference (eg. 'left' or 'right' from an egocentric perspective or from the actor's perspective). Although such results showed that some participants spontaneously took the actor's perspective, which is a mentalizing behaviour and might relate with belief reasoning, this makes us unable to use the accuracy of these attention check questions as a criterion to exclude data. As we were unable to trace back the exact questions we asked, we decided to exclude this criterion in the end.

ANOVA analysis

The raw data were categorised into four outcomes, hit rate, miss, false alarm rate or correct rejection, depending on accuracy and presence of the signal. The SDT *d'* was calculated by subtracting the z score for false alarm rate from the z score for hit rate (see formula 1) and criterion was calculated using the formula (2) listed below (Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999). These values were then submitted to a repeated-measures ANOVA to test if the Context, Participant's belief, and the Agent's belief can bias the judgments of the participant, and thus could provide evidence for spontaneous ToM under different social context in this task.

$$d' = Z_{\text{hit}} - Z_{\text{false alarm}} \tag{1}$$

$$c = -0.5 * (Z_{\text{hit}} + Z_{\text{false alarm}}) \tag{2}$$

After calculating participant's perceptual sensitivity (*d'*) and decision criteria (*c*) under each condition, all dependent variables (*d', c*, accuracy and reaction time) were first submitted to a 2 × 2 × 2 repeated measures analysis of variance (ANOVA) with Context (interactive; non-interactive), Participant's belief (Dash present, Dash absent) and Agent's belief (Dash present, Dash absent) as within-subject factors. Trials with reaction time (RT) below 150 ms or above the three standard deviations from the participant's mean RT were excluded from further analysis.

Bayesian Hierarchical Model

Corresponding to the analyses in Chapter 2, we then adopted a Bayesian approach which is less influenced by extreme *d'* and *criteria* values. The Bayesian approach allows us to obtain the posterior distributions of regression coefficients encoding the influence of our 2 x 2 x 2 factorial design on group-level sensitivity and criterion parameters. Unlike the hierarchical model we used in Chapter 2, the SDT model was nested inside a regression model that encoded the three factors of our experimental design (Context × Participant's beliefs × Agent's beliefs), plus their interactions. Thus each subject's *d'* parameter was specified as:

$$d' = d'_{\text{base}} + \beta_{\text{c}} * I_{\text{c}} + \beta_{\text{pb}} * I_{\text{pb}} + \beta_{\text{ab}} * I_{\text{ab}} + \beta_{\text{cp\_i}} * I_{\text{c}} * I_{\text{pb}} + \beta_{\text{ca\_i}} * I_{\text{c}} * I_{\text{ab}} + \beta_{\text{pa\_i}} * I_{\text{pb}} * I_{\text{ab}} + \beta_{\text{cpa\_i}} * I_{\text{c}} * I_{\text{pb}} * I_{\text{ab}}$$

where I_{c} is an indicator variable that is equal to 1 when the context is interactive and -1 when the context is non-interactive, and the I_{pb} and I_{ab} are equal to 1 when the participant/actor is holding a target present belief and -1 otherwise. The β variables are the regression coefficients. β_{c} encoded the effects from the social contexts, and β_{pb} , β_{ab} represents the effects of the participant's beliefs and the agent's beliefs. $\beta_{\text{cp\_i}}$, $\beta_{\text{ca\_i}}$, and $\beta_{\text{pa\_i}}$ are regression coefficients encoding the effects of two-way interactions between the social context, the participants' beliefs and the agent's beliefs. Lastly, the $\beta_{\text{cpa\_i}}$ regression coefficient represents the three-way interaction. Uninformative (high variance) priors on these influences on *d'* were specified as follows (after JAGS convention, variances are written as precisions or the reciprocal of variance):

$$d'_{\text{base}} \sim N\,(0, 0.001)$$

$$\beta_{\text{c}} \sim (0, 0.001)$$

$$\beta_{\text{pb}} \sim (0, 0.001)$$

$$\beta_{\text{ab}} \sim (0, 0.001)$$

$$\beta_{\text{cp\_i}} \sim (0, 0.001)$$

$$\beta_{\text{ca\_i}} \sim (0, 0.001)$$

$$\beta_{\text{pa\_i}} \sim (0, 0.001)$$

$$\beta_{\text{cpa\_i}} \sim (0, 0.001)$$

Analogous models and parameter estimation were also applied for the criterion, *c*. Markov Chain Monte Carlo (MCMC) implemented in JAGS in R was used to draw samples from the posterior distributions. When calling JAGS, we implemented 2000 adaptation steps, 5000 burn-in samples and 50000 effective samples. For each parameter we run 3 chains and convergence of all chains was assessed both visually and using Gelman & Rubin's potential scale-reduction statistic R for all parameters (Gelman, & Rubin, 1992). Our average R was 1.00 with a maximum value of 1.04, indicating good convergence.

The posterior distributions of each parameter returned by JAGS were then directly used for Bayesian inference. Similar to Chapter 2, for each indicator variable, we calculated the probability that the coefficient is smaller than zero. For example, P (β_{c}) for *d'* stands for the probability that regression coefficients for the Context factor on *d'* is smaller than zero. To distinguish these probabilities from classical P-values, we denote them as P_{θ}.

3.3 Results

ANOVA analysis

According to the data exclusion criteria, 7 participants were removed from the ANOVA analysis in total (3 due to insufficient overall performance < 0.55 and 4 for response bias). We first would like to replicate our findings from Chapter 2, therefore the data for perceptual sensitivity (*d'*), decision criterion (*c*), reaction time (RT) and accuracy in the non-interactive condition only were submitted to a 2 × 2 repeated-measures ANOVA, with Participant's belief (Dash present, Dash absent) and Agent's belief (Dash present, Dash absent) as within-subject factors. In the next step, we added Context (interactive, non-interactive) as another within-subject factor and conducted a three-way repeated-measures ANOVA on all measurements. For each index in the following paragraphs, we shall first present the result from the non-interactive condition, then that including the Context factor.

Perceptual sensitivity (d')

We analysed participants' perceptual sensitivities from the non-interactive condition. A 2 × 2 repeated-measures ANOVA, with the Participant's belief (Dash present, Dash absent) and the Agent's belief (Dash present, Dash absent) as within-subject factors was applied to the data. Similar to the results in Chapter 2, no main effect or interaction is significant.

We then applied a three-way repeated measurement ANOVA including the Context (interactive; non-interactive) factor. The social context has no significant influence on participants' perceptual sensitivities, nor does it interact with the Participant's or the Agent's beliefs. The averaged *d'* value was 1.33 across all conditions, suggesting that participants were able to discriminate signal from noise better than by chance (d' = 0 suggesting $Z_{hit} = Z_{fa}$, see Abdi, 2007).

Decision criterion

In the previous study (Chapter 2), we discovered that in the minimally social environment when no interaction was present, participant's decision criteria were slightly more liberal when they learned that the actor believed the target was present. To formally test if our data replicates this finding, we first took out the criteria data from the non-interactive condition and conducted a 2 (Participant's beliefs) × 2 (Agent's beliefs) repeated-measurement ANOVA. Interestingly, results indicated a similar marginal main effect from the Agent's beliefs (F (1, 47) = 3.56, p = .065). When the agent believed the robot was behind the occlude, participants tended to make more liberal decisions. The main effect of the Participant's beliefs was not significant and neither was their interaction (Table 3-1.).

Table 3-1. ANOVA analyses results of decision criteria (C stands for Context, P stands for Participant’s beliefs, A stands for Agent’s beliefs).

Criteria			
	F	*p*	η_p^2
Non-interactive condition			
Participant's beliefs	1.49	0.229	0.031
Agent's beliefs	3.57	0.065	0.071
P x A	0.03	0.870	0.001
All conditions			
Participant's beliefs	**4.93**	**0.031**	**0.095**
Agent's beliefs	2.32	0.134	0.047
Context	**7.20**	**0.010**	**0.133**
P x A	0.48	0.491	0.010
P x C	0.31	0.580	0.007
A x C	1.45	0.234	0.030
P x A x C	0.18	0.677	0.004

We then included the Context factor into analysis and conducted a three-way repeated measurement ANOVA. The context factor has a significant main effect (*F* (1, 47) = 7.20, *p* = .01, η_p^2 = .133, Table 3-1.). In the interactive condition, participants tended to adopt more liberal decision criteria. Participant’s beliefs also have a significant main effect on criteria (*F* (1, 47) = 4.93, *p* = .031, η_p^2 = .095, Table 3-1.), when participants believed Dash was present their criteria was more liberal. No other main effect or interaction reached the significant level (Figure 3-5.). Corresponding to what we observed from the results in Chapter 2, participants’ averaged decision criteria across all conditions was positive, suggesting they adopted stricter-than-ideal criteria during the task (Abdi, 2007).

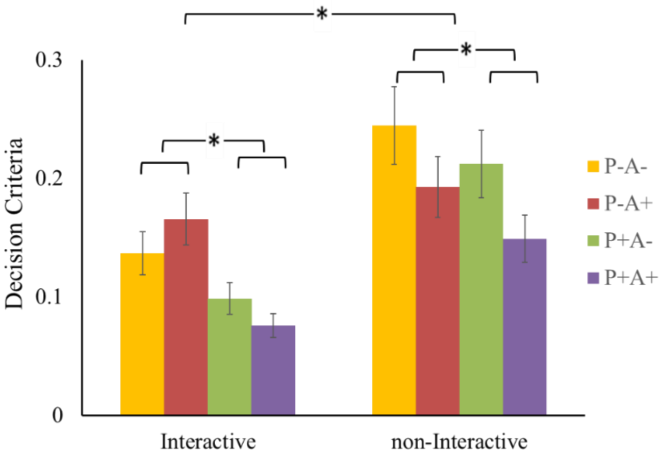


Figure 3-5. Mean of the decision criteria across all conditions.

Accuracy

The mean overall accuracy across all eight conditions is 0.72 suggesting our titrating manipulation is effective to control the task difficulty (to achieve a ~0.7 accuracy in a yes/no paradigm). We then followed the previous procedures and first analysed the accuracy in the non-interactive condition. There was no main effect of the Participant's beliefs or the Agent's beliefs on participants' accuracy. Their interaction was also not significant.

When adding the Context factor, again none of the three factors have main effects on accuracy across different conditions, nor do they have any interactions.

Reaction Time

First, we examined RT in the non-interactive conditions only to test if these replicate results in Chapter 2, where we found a significant main effect from Agent's beliefs. A 2 × 2 repeated-measures ANOVA was applied, with the Participant's beliefs (Dash present; Dash absent) and Agent's beliefs (Dash present; Dash absent) being the within-subject factors. Results revealed there was a significant interaction between the two factors (*F* (1, 47) = 9.727, *p* = .003, η_p^2 = .171), when participants believed Dash was not present, their reaction time was shorter if the agent falsely believed Dash was present. This difference disappeared when participants believed Dash was present. There was also a significant main effect from Participant's beliefs (*F* (s1, 47) = 18.906, *p* < .001, η_p^2 = .287, Figure 3-6.). Participants' reaction time was shorter when they themselves believed the robot was behind the occluder. Unlike the results in Chapter 2, the main effect of the Agent's beliefs was not significant (Table 3-2.).

A three-way repeated measurement ANOVA including Context as the third factor was then applied. Results showed that Participant's beliefs still have a significant main effect, that reaction time was shorter when participants believed Dash was present (*F* (1, 47) = 17.639, *p* < .001, η_p^2 = .273). But this main effect was modulated by the Context factor as the interaction between Participant's beliefs and Context was significant (*F* (1, 47) = 6.009, *p* = .018, η_p^2 = .113). The Context factor also has a significant main effect ((*F* (1, 47) = 18.072, *p* < .001, η_p^2 = .278), that participant's reaction time was significantly shorter in the interactive condition. The interaction among the three factors was also significant ((*F* (1, 47) = 4.173, *p* < .047, η_p^2 = .082)

Table 3-2. ANOVA analyses results of reaction time (C stands for Context, P stands for Participant's beliefs, A stands for Agent's beliefs).

Reaction Time			
	F	*p*	η_p^2
Non-interactive condition			
Participant's beliefs	**18.906**	**<.001**	**0.287**
Agent's beliefs	0.903	0.347	0.019
P x A	**9.727**	**0.003**	**0.171**
All conditions			
Participant's beliefs	**17.639**	**<.001**	**0.273**
Agent's beliefs	3.615	0.063	0.071
Context	**18.072**	**<.001**	**0.278**
P x A	**5.705**	**0.021**	**0.108**
P x C	**6.009**	**0.018**	**0.113**
A x C	0.484	0.490	0.010
P x A x C	**4.173**	**0.047**	**0.082**

To better interpret the results, we considered the P-A- condition in each of the social context as the baseline, and compared RT of the other three conditions with them (Kovács et al., 2010). The P-A- condition was selected as the baseline because in this condition neither the participant nor the agent would believe Dash was present behind the occluder, so RT in this condition should be the longest when the social context is constant. For the non-interactive condition, RT in P–A– is significantly longer than in all other three conditions (P–A+: t (47) = 2.66, *p* =.01; P+A–: t (47) = 5.19, *p* < .001; P+A+: t (47) = 3.99, *p* < .001. In line with Kovács' results, RT in the P-A+ condition is longer compared with that of the P-A- condition, where the only difference is agent belief. We then performed a similar analysis for the interactive condition, and results showed that RT in P–A– was marginally longer than in the P–A+ condition (t (47) = 1.91, *p* =.063), and significantly longer than the P+A+ condition (t (47) = 2.98, *p* = .005). There was no difference between P-A- and P+A-. All the p-values were corrected with Bonferroni correction for multiple comparisons. These results suggest that although the agent's beliefs seem to influence the decision-making time across all contexts, the participant's own belief seems to have a more robust effect in the non-interactive condition.

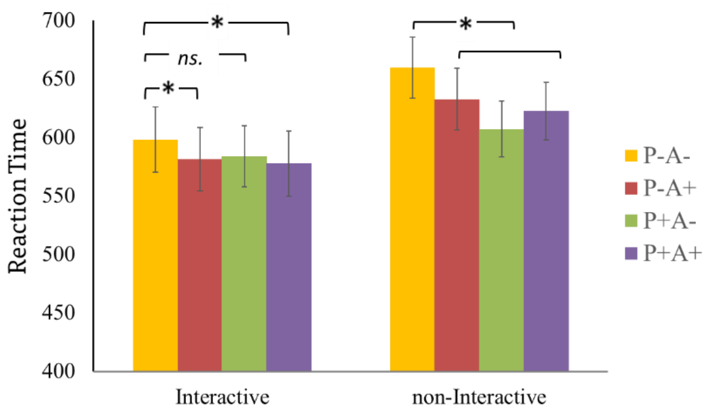


Figure 3-6. Reaction time across all conditions.

Questionnaire results

We also analysed the correlations between the AQ scores, participants' ratings on the 12 post-experiment questions with main indices from the belief task. The differential scores of *d*', criteria, accuracy and RT were first calculated by subtracting the mean scores of the A- conditions from the A+ conditions. Then Pearson correlation analyses were used to test whether any correlational coefficient reached a significant level. However, no significant correlation was reported as significant.

Bayesian Analysis

Since the Bayesian hierarchical model is less influenced by extreme values from the sample, consistent with the procedure in Chapter 2, we also included data with low accuracy or biased report into this part of analyses. Hence, all 55 participants' data were included in the Bayesian model to estimate the parameter values for each factor and the interaction effect.

For *d'*, consistent with our findings from the canonical ANOVA analyses, the Bayesian analysis didn't reveal any larger probabilities for any main effect or interaction effect. However, for the Bayesian model on criteria, the analysis showed large probabilities for participant's belief (P_{Θ} = .99), for the Context factor (P_{Θ} = .95) and a small probability for the interaction between the Context factor and the Agent belief (P_{Θ} = .05), suggesting these factors are very likely to influence participants' decision criteria (Figure 3-7.). It is worth noting that in the canonical ANOVA analysis on criteria, the interaction between Context and the Agent's beliefs was not significant. The reason for the discrepancy of results between these two analysis methods maybe because we include more data in the Bayesian analysis. In general, these results support the hypothesis that when participants held a 'Dash is present' belief, they are more likely to report to see Gabor on Dash. Also, from the Figure 3-8. we can see that in the non-interactive condition, when the agent believed Dash is present, participants also tended to report signals, but such difference did not exist in the interactive context.

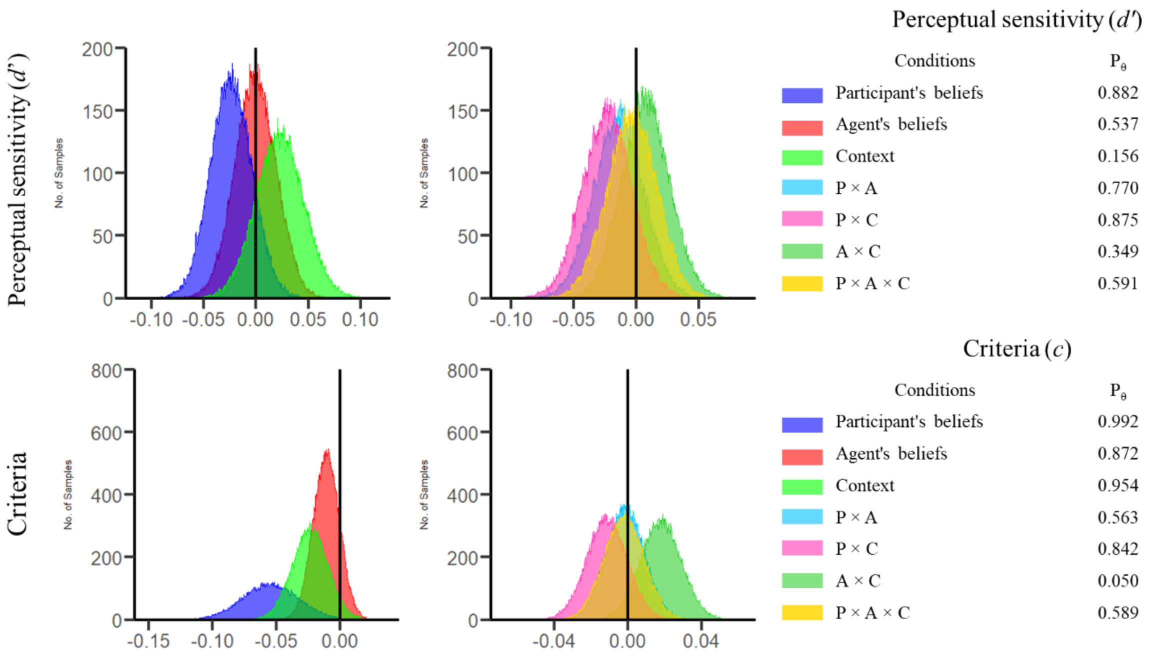


Figure 3-8. The Bayesian Hierarchical Analysis results. The top row of figures displays the results of perceptual sensitivity, and the bottom row of figures are the results of criteria. 'P' stands for 'Participant's belief', 'A' for 'Agent's belief' and 'C' for 'Context'

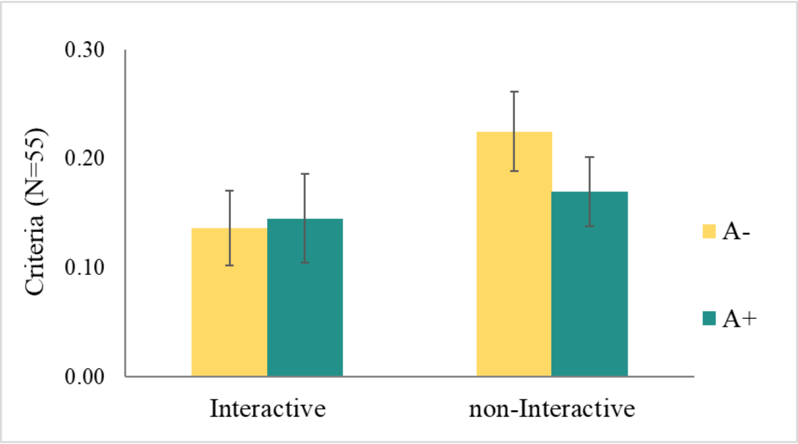


Figure 3-7. The criteria results in the non-interactive condition.

3.4 Discussion

In this study, we aimed to test whether the agent-object interaction has any influence in boosting spontaneous theory-of-mind (ToM). Thus, we created two different social contexts: a more interactive context which has the information of the relationship between the agent and the object and displays goal-directed actions as interactive cues, and a non-interactive condition as a control condition which is parallel with our paradigm in Chapter 2. Based on the findings in previous studies, we predicted that in a more interactive context, participants have more reasons to be involved in the mentalizing process, thus we should be able to observe a larger influence from the agent's beliefs on our perceptual processes. To our surprise, although in the non-interactive context we discovered a weak influence from the agent's beliefs, all effective indices (criteria and reaction time) seem to suggest a smaller effect from the agent's beliefs in the interactive condition. Such results are further confirmed by the Bayesian Hierarchical Model analysis. In the discussion, we shall first provide possible reasons for these counterintuitive results, then discuss the limitation of the current study and possible improvement methods.

3.4.1 Current findings on spontaneous ToM

In the current study, the analysis of the decision-making criteria of the SDT measurement in the non-interactive context suggests a consistent pattern with what we found in Chapter 2. In both studies the main effect of Agent's belief on criteria is marginally significant, even though we tested a large sample size (N = 40 in Chapter 2 and N = 48 in this study), indicating the influence from other's beliefs in the non-interactive social context is a weak effect. Similar result patterns also showed on reaction time: in Chapter 2 we found a significant main effect from the Agent's beliefs, and here there was a significant interaction between Participant's belief and Agent's belief. It is worth noting that the significant interaction effect was driven by longer RT in the P-A- condition compared with the result in Chapter 2, wherein this condition both the participant and the agent believed that Dash was not present. Thus, the difference between the RT results in these two studies is likely to be driven by a stronger influence imposed by participants' own beliefs. The effect of the Participant's beliefs is possibly due to that participants have a stronger motivation to complete the current task, as the videos we used in this study were filmed in the real-life settings and both the actor and the object were real. In that, participants might be more involved when they were watching the videos. In general, the results in Chapter 2 and in the non-interactive context of the current study hinted that when no additional social cues are present, people have a weak tendency to spontaneously represent other's mental state.

When including the Context factor into the analyses, on the contrary, the results are against our hypothesis that a richer social context would boost mentalizing. The more interactive context in the current study did not enlarge the interference effect from other's mental state, instead, it narrowed the difference between the different belief conditions. This result is further supported by the Bayesian Hierarchical Model analyses, which indicated a larger effect of Agent belief in the non-interactive than the interactive condition is highly possible ($P_{\Theta} = .05$). In retrospect, we listed several reasons which may cause such a surprising result.

3.4.2 Possible reasons for the lack of spontaneous ToM in the interactive context

Our interactive condition set-up is imperfect in several ways, which may reduce the explanatory power of our current result and generate the opposite result pattern from our hypothesis. In the first place, as we stated in the ‘Result’ session, the attention check questions were not effective for us to screen if participants paid enough attention to the task or not. Two-thirds of the questions asked where the actor was looking, however the answers provided were ‘left’ or ‘right’, which prompted participants to take the actor’s perspective and many participants hence gave wrong answers. As we thus excluded this criterion when screening data, we might include participants who didn’t attend to the videos and thus increased the standard deviation of the sample.

However, results in the non-interactive context suggest the confounding effect from the exclusion of attention-check criteria might be small, as the result patterns on decision criteria and reaction time consistently support the hypothesis that participants were subtly influenced by the agent’s beliefs. We thus believe the main factor influencing the result may be the reaching behaviour we inserted at the end of each video in the interactive condition, which caused the drop both in criteria and RT across all belief conditions in the interactive context. To deliver the critical message that the agent cares about Dash’s location in the interactive context, when filming each of the interactive videos, we asked the actor to reach the wooden occluder as if she was going to pull it down. We adopted this manipulation from the previous anticipatory looking paradigm (Senju et al., 2007; Schneider et al., 2012). However, such a goal-directed movement may mislead participants to think that the actor always believed Dash was behind the occluder, thus masked the critical belief information we manipulated beforehand. This alternative account is suggested by the difference in criteria between the non-interactive context and the interactive context, as participants became more liberal in general and inclined to report signal. The reaching behaviour may also serve as an attentional cue, as after several trials participants may learn that the testing phase was right after this movement, thus became more prepared for the keypress, which might shorten their reaction time.

3.4.3 Limitations and future direction

As aforementioned in the previous paragraphs, there are several limitations of the current study which may cause our interactive context set-up to be less effective. The most obvious flaw is the attention questions we asked, which caused misunderstandings when participants were providing answers. Future studies should guarantee the questions and answers were straightforward. Future studies may also consider using more objective indices such as eye-gaze to monitor if participants paid enough attention to the videos. Such measurements will also allow us to see whether spontaneous ToM is related to the length of time participants spending on watching the video or not.

Another imperfection is also related to the interactive context we set up. The goal-directed reaching behaviour in this context may not only deliver the social information of the actor's intention as we desired, but also mask the actor's belief information and become a visual cue indicating the response window. To remove the latter two confounding effects, we could adopt the approach used in the anticipatory looking paradigm to make the actor reach the target after participants have made a keypress. In this way, participants' judgments would less likely be influenced by this goal-directed behaviour right before their keypress, but participants will still have reason to represent the agent's beliefs as she will look for the target. Thus, the interaction between the actor and the target remains but participants would not be able to use such cues to predict the responding window.

Another factor which can be better controlled is the appearance of the robot. To simplify the current design, in this research we only used the robot with colored stickers for the interactive condition and the one with white stickers for the noninteractive condition. Such manipulation might also be the cause for the RT difference between these two conditions, i.e. the colourful stickers make the robot more salient in the interactive context even though we masked the bottom of the robot in the response stage. Future studies may consider balancing this factor among different participants.

3.5 Conclusion

In this study, we adopted the signal detection methodology used in Chapter 2 and set up a socially interactive and non-interactive context to test the spontaneous mentalizing process. Participants viewed a female actor interacting (or not interacting) with a toy robot, and were required to detect the Gabor pattern on the robot's helmet. In the non-interactive condition, we found similar results patterns with that in Chapter 2. Such consistency of results in these two studies indicates the spontaneous mentalizing process in the minimal social setting is subtle and the interference from other’s mental state is a weak effect. In the interactive condition, on the other hand, we failed to reveal a larger effect from other’s beliefs. The main reason might be the misplacement of the goal-directed reaching behaviour. In general, the current study showed that social context might have a general effect in changing participant’s perceptual process, however whether it would change how other’s beliefs influence our own perception is still an open question.

References

Abdi, H. (2007). Signal Detection Theory (SDT). *Encyclopedia of Measurement and Statistics*, 1–9.

Baldassano, C., Beck, D. M., & Fei-Fei, L. (2017). Human-Object Interactions Are More than the Sum of Their Parts. *Cerebral Cortex (New York, N.Y. : 1991)*, *27*(3), 2276–2288. https://doi.org/10.1093/cercor/bhw077

Baron-cohen, S., Leslie, A., & Frith, U. (1985). The autistic child have a “theory of mind”? *Cognitive Development*, *21*, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome / High-Functioning Autism , Males and Females , Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1).

Catmur, C., & Heyes, C. (2019). Mirroring “meaningful” actions: Sensorimotor learning modulates imitation of goal-directed actions. *Quarterly Journal of Experimental Psychology (2006)*, *72*(2), 322–334. https://doi.org/10.1080/17470218.2017.1344257

David A. Rosenbaum, Kate M. Chapman, Matthias Weigelt, D. J. W., & Wel, R. van der. (2012). Cognition, Action and Objcet Mnipulation. *Psychological Bulletin*, *138*(5), 924–946. https://doi.org/10.1037/a0027839.COGNITION

Ding, X., Gao, Z., & Shen, M. (2017). Two Equals One: Two Human Actions During Social Interaction Are Grouped as One Unit in Working Memory. *Psychological Science*, *28*(9), 1311–1320. https://doi.org/10.1177/0956797617707318

Errante, A., & Fogassi, L. (2019). Parieto-frontal mechanisms underlying observation of complex hand-object manipulation. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-018-36640-5

Gelman, A., Rubin, D. B., Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences Linked references are available on JSTOR for this article : Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. http://doi.org/10.1214/ss/1177011136

Kim, J. G., Biederman, I., & Juan, C. H. (2011). The benefit of object interactions arises in the lateral occipital cortex independent of attentional modulation from the intraparietal sulcus: A transcranial magnetic stimulation study. *Journal of Neuroscience*, *31*(22), 8320–8324. https://doi.org/10.1523/JNEUROSCI.6450-10.2011

Klapp, S. T., & Jagacinski, R. J. (2011). Gestalt Principles in the Control of Motor Action. *Psychological Bulletin*, *137*(3), 443–462. https://doi.org/10.1037/a0022361

Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, *330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Phillips, J., Ong, D. C., Surtees, a. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A Second Look at Automatic Theory of Mind: Reconsidering Kovacs, Teglas, and Endress (2010). *Psychological Science*, 1–15. https://doi.org/10.1177/0956797614558717

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*(3), 433–438. https://doi.org/10.1037/a0025458

Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., … Csibra, G. (2010). Absence of spontaneous action anticipation by false belief attribution in children with autism spectrum disorder. *Development and Psychopathology*, *22*(2), 353–360. https://doi.org/10.1017/S0954579410000106

Seow, T., & Fleming, S. M. (2019). Perceptual sensitivity is modulated by what others can see. *Attention, Perception, and Psychophysics*, *81*(6), 1979–1990. https://doi.org/10.3758/s13414-019-01724-5

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

https://doi.org/10.3758/BF03207704

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120. https://doi.org/10.3758/BF03202828

Yin, J., Xu, H., Duan, J., & Shen, M. (2018). Object-Based Attention on Social Units: Visual Selection of Hands Performing a Social Interaction. *Psychological Science*, *29*(7), 1040–1048. https://doi.org/10.1177/0956797617749636

Yu, C., & Smith, L. B. (2016). The Social Origins of Sustained Attention in One-Year-Old Human Infants. *Current Biology*, *26*(9), 1235–1240. https://doi.org/10.1016/j.cub.2016.03.026

Chapter 4. Taking the perspectives of many people -- humanization matters

4.1 Abstract

In a busy space, people encounter many other people with different viewpoints, but classic studies of visual perspective taking (VPT) examine only one agent at a time. This paper explores the issue of selectivity in VPT when different people are available to interact with. We consider the hypothesis that humanisation impacts on VPT in four studies using virtual reality methods. In Experiment 1, we tested the humanization hypothesis in an intergroup context and found that participants were more accurate when taking an ingroup agent's perspective. In Experiment 2, we contrast a moving agent against a statue agent, and revealed that participants were faster when reporting items facing the moving agent. All results support the claim that humanisation alters the propensity to engage in VPT in rich social contexts.

4.2 Introduction

Walking into a busy shop, the shopper might encounter a number of other figures such as a shop worker, a friend and a shop mannequin, who all relate differently to us and all have different visual perspectives on the scene. In Chapter 3, we investigated whether the agent-object interaction has any influence on mentalizing. In this study, we would like to ask whether the self-agent interaction can influence mentalizing. Specifically, this study questioned if different relationships between self and the agent can change how people engage in visual perspective taking (VPT), when they encounter multiple different agents with different social characteristics. In particular, we aim to examine the tension between claims that we automatically consider the visual perspective of people we encounter (Samson et al., 2010) and the suggestion that not all people we encounter are fully humanised (Haslam, 2006) , together with the real-world observation that we often meet more than one person at a time.

Visual perspective taking is the process of determining if another person can see an object and what the object looks like to that person (Flavell, 1977). Many cognitive studies over the last decade have suggested that at least some forms of VPT are automatic and occur without top-down control (Samson et al., 2010; Furlanetto, Becchio, Samson, Apperly, 2015;

Surtees, Apperly, & Samson, 2016a; Surtees, Samson, & Apperly, 2016b; Elekes, Varga, & Király, 2016, 2017; Freundlieb, Kovács, & Sebanz, 2016, 2018). Samson (2010) found an 'altercentric intrusion effect' when participants were asked to report the number of discs on the walls and a human agent could see a different number of discs than them. Recently, Ward and colleagues found that recognition of rotated letters was easier when the letters were oriented towards another person (Ward et al., 2019). Similarly, lexical decisions on rotated words are easier when the words are oriented to another person (Freundlieb, Sebanz, & Kovács, 2017; Freundlieb et al., 2018). Thus, in different contexts, the presence of another person can either interfere with (Samson's task) or facilitate (social mental rotation task) participant's judgments of what they themselves can see. These studies have typically used rapid reaction time measures in tightly controlled environments and suggest rapid or even automatic mechanism of processing other's visual perspectives.

This contrasts with studies using tasks that give more time for thought. For example, researchers found that adults are more likely to draw an E on their own forehead to be readable by another person if the participant feels less powerful (Galinsky, Ku, & Wang, 2005) or if the confederate is from in-group (Vaes, Paladino, & Leyens, 2004). Similarly, young children use more metalizing words when describing in-group members (McLoughlin & Over, 2017), and adults are more likely to attribute secondary emotions to in-group members (Demoulin et al., 2009). Finally, people will spontaneously take the perspective of a human more than a robot in a single-trial online study (Zhao, Cusimano, & Malle, 2016). These results can be summarised in terms of humanisation (Gray, Gray, & Wegner, 2007), that is, the theory that people do not attribute as many human abilities, including emotion and perspective taking, to robots and outgroup members compared to in-group members. However, this has rarely been tested in cognitive VPT tasks.

Thus, in this study we would like to manipulate the 'self-other' interaction in the mentalizing triangle, and ask how VPT works when people interact with agents with different level of humanisation in a simple cognitive VPT task. To make a direct comparison of VPT performance with different agents, this present study thus aims at studying VPT in more naturalistic contexts than previous studies with one agent. The majority of previous studies have used a static photo of a neutral person as a stimulus, and give us little information about how perspective-taking works in real-world contexts where there are many people who can move. In the present study, we used virtual reality to test the hypothesis that humanisation acts as a gateway to VPT and that participants selectively take the perspectives of agents who are more humanised, even in rapid response tasks.

Using virtual reality, we implemented a social mental rotation task based on Ward's study (Ward et al., 2019). Participants sat in VR room in front of a table with two agents sitting on the left and right. On each trial, a rotated letter appeared in the centre orienting either to one of the agents or to the participant, and participant must report whether the letter was canonical or mirror-reversed. Since VR allows us to manipulate how agents move and appear, it is thus possible to systematically test whether different levels of humanization change the propensity to engage in VPT. We first conducted a pilot study to select the appropriate experimental stimuli and calculate proper sample size. Then in two experiments, we examined the influence of humanization on VPT from two different levels: an in-group member VS. an out-group member, and a naturally-moving agent VS. a statue agent. In both contrasts, we hypothesize that people are more likely to take the perspective from the agent with a higher level of humanness (e.g. the in-group agent and the moving agent).

4.3 Pilot Study

Participants

29 (18 females, aged 23.5±3.2) right-handed participants took part in this experiment; however, five participants were excluded from further analysis because the post-experiment survey showed that they detected the purpose of our study. This resulted in a final sample size of 24 participants. Participants were recruited from two UCL-associated psychology databases and were required to have the Latin alphabet as the basis of their first language and normal or normal-to-corrected vision. Participants were paid based on a rate of £7.5 per hour.

Materials & VR setup

VR setting was created by in Vizard 5.0 (Worldviz, USA). Participants wore the Oculus Rift DK2 and saw a virtual room where a wooden table was placed with two female agents sitting on the left or right (Figure 1A). Both agents had a European appearance and moved according to the Vizard 'quiet sitting' animation with breathing and small movements but no head turns. We added a small rotation to agents' heads and neck so both agents gazed towards the table as if they were looking at the stimuli. On each trial, one letter was presented in the middle of the table in Arial font from the set F, R, P, G, and Q. After the study was complete, I realised that the letter Q was an ambiguous stimulus without a definite orientation. For example, a normal Q viewed from the participant's location appeared very similar to a reversed Q viewed by a person sitting to the right of the participant. This lead to high error rates for Q and so I decided to remove all the Q trials and present methods and results here only with the remaining four letters (R, F, P, G).

Each block of trials began with the two agents sitting naturally in the virtual room to let participants explore the scene for 5.5 seconds and see who was located where. Then a red dot appeared on the table for 500 ms to draw attention from participants and trials then began. In each trial, a letter appeared in the centre of the table, oriented towards one or the other agent or toward/away from the participant. Participants were instructed to press key ‘J’ if the letter was normal and ‘F’ if it was mirrored as fast as possible. After the keyhit, the letter disappeared and there was an ISI of 900-1100ms before the next trial began. Note that the room and agents all remained visible during the ISI, to maintain the feeling of being in a real location with real people.

The experimental trials are organized in two blocks with a total of 72 trials. The two agent’s sitting positions were balanced between blocks. In each block, there were 32 trials where letters would face towards one of the agents, and two trials where letters would face to the participants and another two face away from the participants. We added these four distracting trials make it harder for participants to detect the purpose of the study. Between the two blocks, there was a break of at least 1.5s then participants could press the ‘space’ when they were ready to resume.

Design

The experiment is composed of two parts. In the first part, participants complete a dot-counting task as a group manipulation (Howard & Rothbart, 1980). The dot-counting task has ten trials, and on each trial 50-250 white dots would present on a black screen for 2 seconds then participants were asked to estimate the number of dots by inputing a number between 50-250. After ten trials, the script would inform the participant that she is an ‘over-estimator’ or an ‘under-estimator’, and participants were then instructed to report how much they identify with the results by merging a ring representing ‘self’ with the rings of ‘over-estimator’ or ‘under-estimator’ (Figure 4-1). Participants then received a colourful sticker to remind their roles, with a red sticker stands for ‘over-estimators’ and a blue sticker for ‘under-estimators’. However, the role assigned to each participant was fully randomised. All the stimuli were presented by using the Cogent Toolbox (http://www.vislab.ucl.ac.uk/cogent.php) for Matlab 2016b.

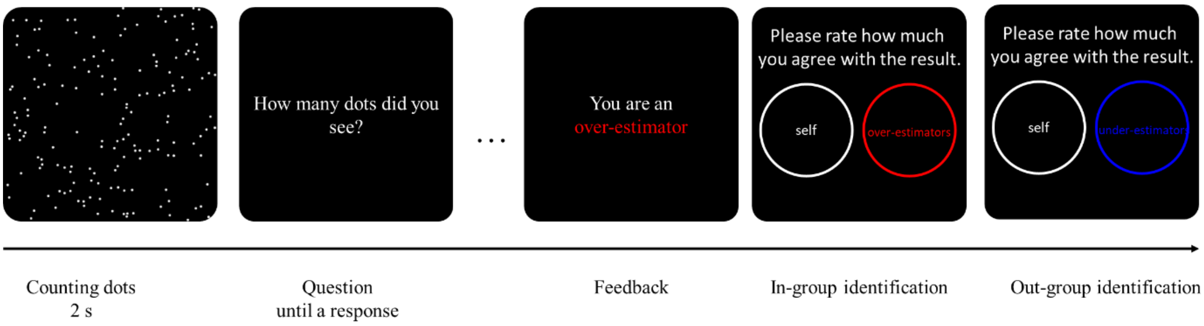

How many dots did you see?
...
You are an over-estimator
Please rate how much you agree with the result.
self
over-estimators
Please rate how much you agree with the result.
self
under-estimator
Counting dots
2 s
Question
until a response
Feedback
In-group identification
Out-group identification

Figure 4-1. The procedure of the dot-counting task.

Then participants were invited to the letter-verification task. To create the in-group and the out-group agent, we selected two female avatars with identical European appearance from the Live Characters in the Vizard 5.0 software. Importantly, participants were informed that the movements of the two agents were recorded from two real participants who also completed the dot-counting task on a previous day, with one avatar’s movements came from an over-estimator and the other from an under-estimator. To make their roles easy to remember to the participants, one of the agents would wear a red sticker on her arm, indicating she was an over-estimator, and the other would wear a blue sticker. The roles of the two agents were balanced across all participants.

Procedure

Participants gave written consent to take part. Their first task was a dot-counting task, used to set up the in-group/out-group manipulation. On each trial, participants saw a black screen with 50-250 white dots for 2 seconds and were asked to judge the number of dots. After 10 trials, participants were randomly assigned the role as an “overestimator” or “underestimators” and received either a red or a blue sticker according to their role, and they were asked to place the sticker on their clothes as an indication of their group membership. This task takes about 10 min to finish.

Participants were then introduced to the two agents (named Lucy and Ellie). Participants were told that the agents were real participants who completed the same task on a previous day, and one of them was an over-estimator and the other an under-estimator. Then they were instructed to put on the HMD and get familiar with the room. Participants then accomplished 72 trials of the VPT task separated into two blocks, wherein between they were instructed to take a break. During the experiment, the two agents would wear coloured stickers according to their roles, and their roles were balanced across participants. Their sitting positions were balanced across blocks. The whole study took about 20 minutes to finish.

Results

A 2×2×2 repeated-measurement ANOVA was applied to analyse accuracy for all trials and reaction time (RT) for correct trials after excluding extreme values (± 3 SDs), with letter-direction (left; right), letter-type (canonical; mirror-inverted) and agent (ingroup, outgroup) as within-subjects factors. For accuracy, no significant main effect or interactions were found related to the agent factor. There was a significant main effect for letter-direction, with more accurate responses for letters towards the right (F (1,23) = 4.81, p = .039, η_p^2 = 0.17, see Figure 4-2.B). As all our participants were right-handed, this result may reflect an effect of handedness.

For RT, a significant main effect was observed for letter-type (F (1, 23) = 26.52, p < .001, η_p^2 = 0.57), with canonical letters being processed more quickly than its mirrored versions. A significant interaction was found for letter-type × agent (F (1, 23) = 7.36, p = .012, η_p^2 = 0.24), when canonical letters are towards an ingroup member, participants made faster responses (t = -2.62, p = .015), but this was not the case for mirror-inverted letters (t = 0.86, p = .400) (Figure 4-2C).

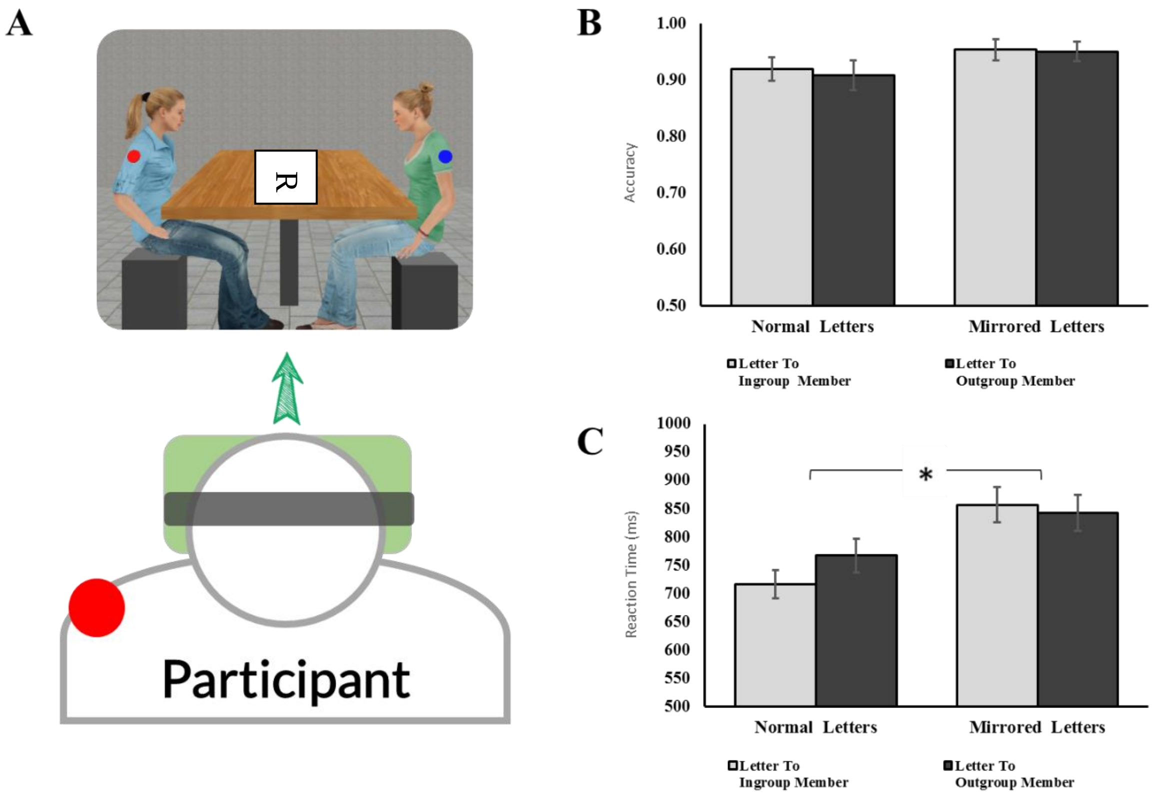


Figure 4-2. The design and results of the pilot study. An illustration of the VR setting in the Pilot study (A); Results on accuracy of recognising letters F, R, P and G (B); Results on reaction time (ms) of recognising letters F, R, P and G (C). The frame of the letter ‘R’ and the white background is added here to make the letter more salient on the figure. It was not present in the experiment.

Discussion

Results from our pilot study show that the current paradigm is effective to investigate perspective selection in multi-perspective scenarios. Being exposed simultaneously to perspectives both from an ingroup and an outgroup member, participants were able to recognise normal letters faster when the letter oriented towards an ingroup member, indicating they have a stronger propensity to take the in-group member's perspective. These results also showed that our two-person mental rotation task is effective to test perspective selection. We then used the results from the pilot study to calculate the proper sample size for our formal study in G*power 3.1.

4.4 Experiment 1

Following our pilot study, Experiment 1 intended to compare perspective-taking of an ingroup and an outgroup agent with a more robust measure. We adopted the same minimal group manipulation in the pilot study to give participants the sense of membership. The hypothesis is that when letters are facing towards an ingroup agent, as participants have a stronger propensity to take their perspective, they should be either quicker or more accurate to recognise the letters.

Participants

We calculated the sample size using G*power 3.0 based on the results of our pilot study. To achieve an power of 0.8 on a .05 significant level, 36 right-handed participants were recruited in this experiment (23 females, mean age = 25.4, *SD* = 4.90). Participants were required to have the Latin alphabet as the basis of their first language and normal or normal-to-corrected vision. Payment and recruitment details are the same as in the previous two experiments.

Materials & VR setup

The VR setting was created by in Vizard 5.0 (Worldviz, USA). Participants wore the Oculus Rift DK2 and saw a virtual room where a wooden table was placed with two female agents sitting on the left or right (Figure 4-3.A). Both agents had a European appearance and moved according to the Vizard 'quiet sitting' animation with breathing and small movements; the agent's head was oriented so that they were looking towards the centre of the table where the stimuli appeared. We used four asymmetric letters (F, R, P and G) in Times New Roman font as our stimuli.

In the beginning of each block, the two agents sat naturally in the virtual room by the table for 5.5 seconds to let participants adapt to the scene and get familiar with the positions of the in-group and out-group agents. A red dot then appeared at the centre of the table for 500 ms indicating the upcoming letters. On each trial, a letter appeared in the centre of the table, orienting towards one of the agents or towards the participant. Participants were instructed to press key ‘J’ if the letter was canonical and ‘F’ if it was mirror-reversed. After the keyhit, the letter disappeared and there was an inter-stimulus interval of 900-1100 ms before the next trial began. The room and agents all remained visible during the inter-stimulus interval, to maintain the feeling of being in a real location with real people.

Each participant completed two blocks with 48 trials in each of them and a short break between. In each block, there were 16 trials the letters facing towards the in-group agent, 16 trials towards the out-group agent and another 16 toward the participant. By using this manipulation, we endow equal importance to both the agents’ and the participant’s egocentric perspectives. Letter-type (canonical or mirror-reversed) was balanced within each direction. Trials were presented in random order.

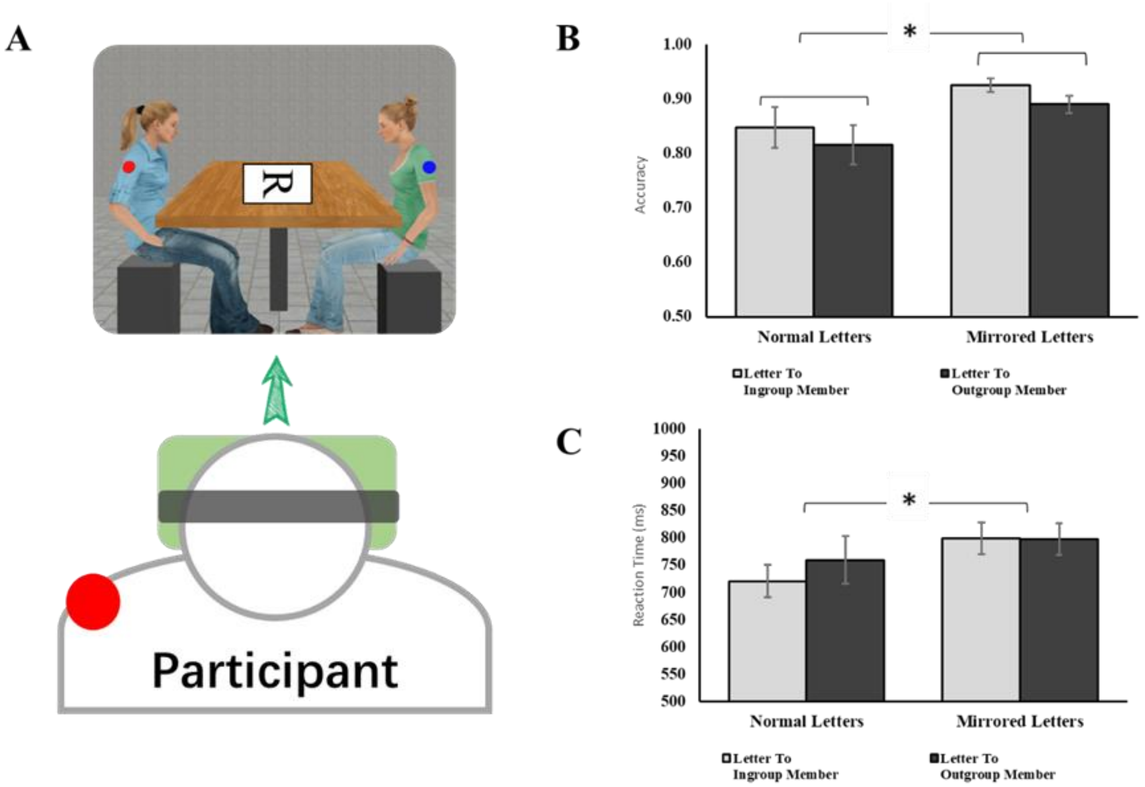


Figure 4-3. The VR setting and results of Experiment 1. (A) The VR environment participants see from the headset, the frame and white background of the letter is for highlighting the letter in the picture, and were not present in the task; (B) Results of accuracy of Experiment 3; (C) Results of reaction time of Experiment 3. The frame of the letter ‘R’ and the white background is added here to make the letter more salient on the figure. It was not present in the experiment.

Procedure

Participants gave written informed consent to take part. The experiment procedure is identical with that in the pilot study, except two changes. First, I only included letters 'R, F, G, P' to avoid ambiguity. Second, to endow equal importance to participant's egocentric and the agents' perspectives, equal numbers of trials were adopted for conditions where letters facing towards each agent or the participant. Participants first underwent the dot-counting task to be assigned the role of an over-estimator or an under-estimator. Then they entered the VR room to start the social mental rotation task with the two agents. In total, participants completed the 96 trials of the social mental rotation task, after which they need to fill in a questionnaire about their strategies and their attitude towards the two agents. The whole study took about 20 minutes to finish (Figure 4-3.A).

Results

A 2×2 repeated-measurement ANOVA was used to analyze accuracy for all trials and reaction time for correct trials after excluding extreme values (±3 SDs), with letter-type (canonical; mirrored-reversed) and agent-group (in-group, out-group) as within-subjects factors. For accuracy, congruent to our hypothesis, a significant main effect was found for agent (F (1, 35) = 8.24, p = .007, η_p^2 = .19). Participants performed better when the letters were facing towards the in-group agent. There was also a significant main effect for letter-type (F (1, 35) = 4.23, p = .047, η_p^2 = .11), with mirror-reversed letters processed better than canonical letters (Figure 4-3.B). For RT, a significant main effect was observed for letter-type (F (1, 35) = 5.48, p = .025, η_p^2 = .14), with canonical letters processed more quickly than mirror-reversed versions (Figure 4-3.C). No other main effect or interaction reached significant level.

4.5 Experiment 2

Experiment 1 showed that, when participants encounter two people with conflicting perspectives, they prefer to take the perspective of the in-group member and their task performance improves for items oriented towards the in-group member. This is the first study to examine perspective selection in multi-agent perspective taking and demonstrates a humanization effect. For our final study, I wanted to explore the effects of humanization with a more subtle manipulation. Thus, I test if participants prefer to take the perspective of an agent who moves like a human, compared to one who is rigid like a statue.

Participants

36 right-handed participants were recruited from two UCL-associated psychology databases (25 females, Mean age = 21.9, *SD* = 3.61). Our requirements and payments for the participants remained the same as in the pilot study.

VR stimuli

The VR setup here was closely modelled on Expt. 1, with the following changes. I did not use the minimal group manipulation, but instead contrast a moving virtual agent who performs natural human actions with one who is rigid like a statute. To avoid having agents switch between ‘moving’ and ‘non-moving’ roles, I used two female agents in blocks 1 and 2, and then two male agents in blocks 3 and 4. The position of the moving agent (left or right) was balanced across the two halves and which agents moved were balanced across participants. The moving agent performed natural seated actions, such as moving the head and torso and looking around, while the still agent was like a statue with no motion. Both agents had a neutral facial expression.

Each block of the task began with a 5.5 seconds familiarization period where participants could look around the virtual space. During this time, the moving agent showed some large movements (turning her head, shifting posture) while the static agent remained rigid. When the trials were about to begin, the moving agent oriented his/her head and body towards the table and showed only the small ‘quiet sitting’ movements, as used in Expt. 3. Again, the static agent remained rigid.

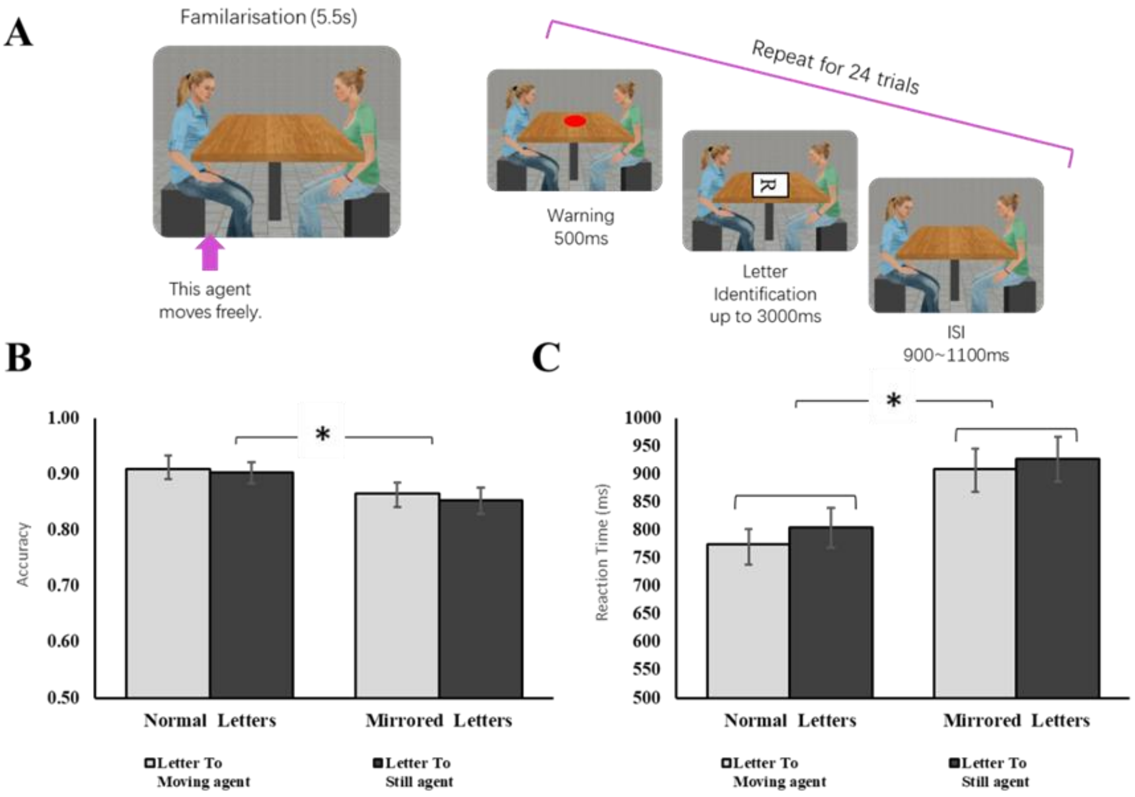


Figure 4-4. The VR setting and results of Experiment 2. (A) The VR environment participants see from the headset and the timeline for a single trial; (B) Results of accuracy of Experiment 4; (C) Results of reaction time of Experiment 4. The frame of the letter 'R' and the white background is added here to make the letter more salient on the figure. It was not present in the experiment.

Procedure

Participants gave written consent and had time to get familiar with the VR before the task began. They completed four blocks of trials with 24 trials per block. As before, each block began with a 5.5s familiarization period where one agent moved and the other was static, followed by the trials (Figure 4-4.A). In each trial, one letter appeared in the centre of the table and participants had to judge if it was canonical or mirror-reversed. In each block, there were eight trials where letters are facing towards the participant, eight towards left and eight towards the right. Letter type was all balanced within each direction and all trials were presented in random order. After the computer-based task, participants were asked to fill out two short questionnaires in which we checked their preference for the two agents and their personalities.

Results

A 2×2 repeated-measurement ANOVA was applied to analyze accuracy for all trials and reaction time for correct trials after excluding extreme values (±3 SDs), with letter-type (canonical; mirror-reversed) and agent (moving, still) as within-subjects factors.

For accuracy, there is a significant main effect for letter-type (*F* (1, 35) = 4.32, *p* = .045, η_p^2 = .11), with participants performed better on canonical letters than on mirror-reversed versions (Figure 4-4.B). No main effect or interaction was found related to agent.

For RT, we found a significant main effect for letter-type, *F* (1, 35) = 45.37, *p* < .001, η_p^2 = .57, and normal letters were processed more quickly than its mirrored versions. Importantly, we observed a significant main effect for agent (*F* (1, 35) = 4.23, *p* = .047, η_p^2 = .11): participants responded quicker when letters were presented to the moving agent. No other main effect or interaction effect reached significant (Figure 4-4.C).

4.6 General Discussion

The aim of this study was to determine if the 'self-other' interaction in the mentalizing triangle matters for VPT, by examining whether people would have different propensity to engage VPT with agents on different humanization levels. Specifically, we modified a new cognitive VPT task to adapt it into contexts with multiple agents. Using virtual reality, we were able to present our moving stimuli in a 3D format with a context that remains constant over all the trials, giving greater ecological validity than typical lab studies. We used a two-person social mental rotation task, and two different manipulations of humanization – an in-group/out-group manipulation and a human/robot-statue manipulation. In all cases, results showed stronger VPT effects for the in-group and human agents, compared with the out-group, robot or statue agents. These results indicate that our initial perception of another agent as human or not plays a critical role in determining our propensity to take the perspective of the other agent.

These results add to our understanding of how perspective taking processes relate to other aspects of cognition. Previous cognitive studies of VPT have emphasised the rapid, even automatic nature of this process (Samson et al., 2010; Furlanetto, Becchio, Samson, & Apperly, 2016; Surtees & Apperly, 2012; Michael, Wolf, Letesson, Butterfill, Skewes, & Hohwy, 2018), or have tried to show modulation of VPT by adding dual tasks or changing motivation (Cane, Ferguson, & Apperly, 2017; Bukowski & Samson, 2016; Todd, Simpson, & Cameron, 2019). Such studies suggest that VPT is relatively impervious to these manipulations. In contrast, our data show that changes in the perception of the agent can change the propensity to engage in VPT. That is, we believe that our data can be understood in terms of two consecutive cognitive processes. First, participants must humanise one agent in the scene, based on a range of criteria. When a human agent is detected, participants can then engage in the process of taking that person's perspective. In the present study, such effect manifests by enhancing what one agent can see when a conflicting perspective co-existed. Overall, we suggest that the humanisation process can act as a gateway to perspective taking, which could then proceed rapidly and spontaneously.

Here, we used several different manipulations of the humanness of our computer-generated agents. We controlled top-down information about the agent with a well-established minimal group manipulation in Experiment 1, however in Experiment 2, we controlled perceptual information about the agent in terms of its appearance as a robot or its movement patterns (natural v. statue). All of these changes had the same impact: decreasing an agent's humanness reduced the propensity of participants to take the agent's perspective. Thus, we interpret all three studies under the general framework of humanisation and dehumanisation, whereby people tend to categorize others into in-group/out-group and affords out-group individuals less positive human essence compared with in-group individuals (Loughnan & Haslam, 2007; Boccato, Cortes, Demoulin, & Leyens, 2007). It seems that robots and statues are treated similarly to outgroup members in our studies, and our results show how processes of humanisation are critical to even basic aspects of social cognition like VPT.

Our current study manipulated the humanness levels of different virtual agents, however, such virtual agents might still not be fully humanised. In real life, we interact with our close friends, colleague, strangers on the street and etc. We sense different 'self-other' relationship with them, and very likely we tend to humanise them differently. Such difference in humanization would further alter the way we process their social information such as their perspectives. We believe future studies on perspective selection might be able to use real people (Freundlieb et al., 2016, 2018) and thus reveal more convincing evidence. Another point worth noting is that humanisation has an impact on different indices in these two experiments, with participants responded more accurately in the intergroup context, however responded faster when exposed to moving and statue agents. We consider this discrepancy may due to a speed-accuracy trade-off, as results from our pilot study of the intergroup setting did reveal faster responses when canonical letters are oriented towards the in-group agent.

In this study, we discovered some interesting findings showing that participants either responded faster or better to items facing towards a more humanised agent. But with the current experiment design it is hard for us to explain why people showed this effect. A number of studies on VPT have shown that visuospatial perspective taking is related to the angular disparity between the participant and the agent (Michelon & Zacks, 2006; Surtees, Apperly & Samson, 2013; Kessler & Thomson, 2010). For example, Surtees, Apperly & Samson (2013) revealed that perspective judgments became more difficult in relevant to the angular disparities between self and other. A possible explanation for our results here then would be that participants may spontaneously posit themselves closer to the more humanised agents, thus their performance on recognizing letters was better. In the following chapter, I will use fNIRS together with VR to further explore the mechanism of this result.

This study focused on an ecological question by asking how people select perspective in real life, considering we often encounter multiple perspectives at the same time. Results emphasize the modulation role of humanisation and suggest our social cognitive capacities might be target-specific. Such results have significant implication for real-life studies as we may humanise other people differently according to the ‘self-other’ relationship. With the digitization of modern society, we believe such results are also valuable for investigating how people interact with various virtual agents, robots, or avatars of our close others in the future. Industrial designers may also consider the way to humanize such agents in order to promote social interaction in the digital world. Future studies may consider testing such findings in real the world, and measuring how such propensity to engage in perspective taking varies in clinical populations.

References

Boccato, G., Cortes, B. P., Demoulin, S., & Leyens, J. P. (2007). The automaticity of infra-humanization. *European Journal of Social Psychology*, *37*(5), 987-999.

Bukowski, H., & Samson, D. (2016). Can emotions influence level-1 visual perspective taking? *Cognitive Neuroscience*, *7*(1–4), 182–191.

Cane, J. E., Ferguson, H. J., & Apperly, I. A. (2017). Using perspective to resolve reference: The impact of cognitive load and motivation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(4), 591–610.

Demoulin, S., Cortes, B. P., Viki, T. G., Rodriguez, A. P., Rodriguez, R. T., Paladino, M. P., & Leyens, J. P. (2009). The role of in-group identification in infra-humanization. *International Journal of Psychology*, *44*(1), 4–11.

Dumontheil, I., Küster, O., Apperly, I. A., & Blakemore, S. J. (2010). Taking perspective into account in a communicative task. *NeuroImage*, *52*(4), 1574–1583.

Elekes, F., Varga, M., & Király, I. (2016). Evidence for spontaneous level-2 perspective taking in adults. *Consciousness and Cognition*, *41*, 93–103.

Elekes, F., Varga, M., & Király, I. (2017). Level-2 perspectives computed quickly and spontaneously: Evidence from eight- to 9.5-year-old children. *British Journal of Developmental Psychology*, *35*(4), 609–622.

Flavell, J. H. (1977). The development of knowledge about visual perception. In *Nebraska symposium on motivation*. University of Nebraska Press.

Freundlieb, M., Kovács, Á. M., & Sebanz, N. (2016). When do humans spontaneously adopt another’s visuospatial perspective? *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 401–412.

Freundlieb, M., Kovács, Á. M., & Sebanz, N. (2018). Reading Your Mind While You Are Reading—Evidence for Spontaneous Visuospatial Perspective Taking During a Semantic Categorization Task. *Psychological Science*.

Freundlieb, M., Sebanz, N., & Kovács, Á. M. (2017). Out of your sight, out of my mind: Knowledge about another person’s visual access modulates spontaneous visuospatial perspective-taking. *Journal of Experimental Psychology: Human Perception and*

Performance, *43*(6), 1065–1072.

Furlanetto T, Becchio C, Samson D, Apperly I, *. (2015). Altercentric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. *Geologia Tecnica e Ambientale*, *19*(3), 55–79.

Furlanetto, T., Becchio, C., Samson, D., & Apperly, I. (2016). Altercentric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(2), 158–163.

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. *Group Processes & Intergroup Relations*, *8*(2), 109-124.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619-619.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264.

Howard, J. W., & Rothbart, M. (1980). Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology*, *38*(2), 301–310.

Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological Science*, *18*(7), 600-606.

Kessler, K., & Thomson, L. A. (2010). The embodied nature of spatial perspective taking: embodied transformation versus sensorimotor interference. *Cognition*, *114*(1), 72-88.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32-38.

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

Loughnan, S., & Haslam, N. (2007). Animals and androids: Implicit associations between social categories and nonhumans. *Psychological Science*, *18*(2), 116-121.

Leyens, J. P., Demoulin, S., Vaes, J., Gaunt, R., & Paladino, M. P. (2007). Infrahumanization: The Wall of Group Differences. *Social Issues and Policy Review, 1*(1), 139–172.

McLoughlin, N., & Over, H. (2017). Young Children Are More Likely to Spontaneously Attribute Mental States to Members of Their Own Group. *Psychological Science*, *28*(10), 1503–1509.

Michael, J., Wolf, T., Letesson, C., Butterfill, S., Skewes, J., & Hohwy, J. (2018). Seeing it both ways: Using a double-cuing task to investigate the role of spatial cuing in Level-1 visual perspective-taking. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(5), 693.

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & psychophysics*, *68*(2), 327-337.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & 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.

Surtees, A. D. R., & Apperly, I. A. (2012). Egocentrism and Automatic Perspective Taking in Children and Adults. *Child Development*, *83*(2), 452–460.

Surtees, A. D. R., Apperly, I. A., & Samson, D. (2013). The use of embodied self-rotation for visual and spatial perspective-taking. *Frontiers in human neuroscience*, *7*, 698.

Surtees, A., Apperly, I., & Samson, D. (2016a). I’ve got your number: Spontaneous perspective-taking in an interactive task. *Cognition, 150,* 43–52.

Surtees, A., Samson, D., & Apperly, I. (2016b). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, *148*, 97–105.

Todd, A. R., Simpson, A. J., & Cameron, C. D. (2019). Time pressure disrupts level-2, but not level-1, visual perspective calculation: A process-dissociation analysis. *Cognition, 189,* 41–54.

Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using Presence Questionnaires in Reality. *Presence: Teleoperators and Virtual Environments, 9*(5), 497–503.

Vaes, J., Paladino, M., & Leyens, J. P. (2004). Perspective taking in an intergroup context and the use of uniquely human emotions: Drawing an E on your forehead. *Revue Internationale de Psychologie Sociale*, *17*(3), 5–26.

Ward, E., Ganis, G., & Bach, P. (2019). Spontaneous Vicarious Perception of the Content of Another's Visual Perspective. *Current Biology*, *29*(5), 874-880.

Zhao, X., Cusimano, C., & Malle, B. F. (2016). Do people spontaneously take a robot's visual perspective?. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 335-342). IEEE.

Chapter 5. Taking the perspective of an in-group member: a Virtual Reality and fNIRS study

5.1 Introduction

The social mental rotation task has been shown to be an effective measurement of spontaneous visuospatial perspective taking (Ward, Ganis, & Bach, 2019; Ward, Gannis, McDonough & Bach, preprint; Ye & Hamilton, in preparation). Ward et al. (2019) found that when an item is oriented towards a person with a different perspective, it can be recognized faster. In our previous studies (see Chapter 4), we outlined that when facing towards two conflicting perspectives from an in-group member and an out-group member, people have a stronger propensity to take the in-group member's perspective. Although the behavioural results are clear-cut, it is still unknown how VPT-related brain areas respond in the one agent social mental rotation task and in the multi-perspective scenario. In this chapter, we would like to investigate the neural mechanism of these spontaneous perspective taking processes using functional near-infrared spectroscopy (fNIRS).

5.1.1 Using fNIRS to investigate brain activity in the social mental rotation task

fNIRS is a non-invasive neuroimaging method which is widely used to enrich our understanding of brain functions in cognitive tasks. Compared with other neuro-imaging methods, such as functional magnetic resonance imaging (fMRI) and electroencephalogram (EEG), the most obvious strength of fNIRS relies on its large tolerance of bodily movements (Pinti et al., 2018). As social psychological experiments place great emphasis on ecological validity, recently fNIRS has become an ideal measurement to uncover the cognitive mechanisms behind social interactions.

Unlike fMRI and EEG, fNIRS is an optical neuroimaging technique. It can measure brain tissue concentration changes of oxygenated (HbO_2) and deoxygenated (HbR) hemoglobin in a specific brain region when people are undergoing cognitive tasks (Pinti et al., 2018). This aim is achieved by shining near-infrared light into the scalp and calculating the light attenuation rate on different wavelengths according to the Beer-Lambert law. In addition to its tolerance to motion artifects, fNIRS also has a good trade-off between spatial (usually 2-3 cm) and temporal (up to 10Hz) resolution compared with fMRI and EEG. Besides, as a harmless and portable instrument, it allows the application on atypical population or measurement in real-life scenarios. Thus, it becomes a preferable choice when participants' safety and comfort are a priority in a study.

In the current study, we would like to investigate the neural mechanism of 1) spontaneous VPT in the one-agent social mental rotation task and 2) perspective selection in the multi-perspective scenario. However, we would like to achieve these while maintaining the feeling of being in a real place with real people. Thus, using fNIRS together with the VR setting offers an ideal option. Since the fNIRS instrument is flexible in the placement of the optical fibres (or optodes), we thus designed our optodes layout and adapted the VR headset to make it suitable for testing with fNIRS (see Figure 5-1). Importantly, we first identified a number of brain regions related to mentalizing process according to previous studies.

5.1.2 The mentalizing neural network

Over recent decades, studies have identified a number of brain regions actively involved in switching our perspective and considering other's mental content. Using the positron emission tomography (PET) approach, Fletcher et al. (1995) first revealed that processing social stories activates a number of brain regions including the bilateral temporal poles, the left superior temporal gyrus and the medial prefrontal cortex. Critically, the prefrontal cortices are specifically involved in processing theory-of-mind stories when compared with physical stories. Later, an fMRI study compared conditions where ToM stories were presented through distinct perceptual modalities (verbal; non-verbal), and revealed consistent results with Fletcher's findings (Gallagher, Happé, Brunswick, & Fletcher, 2000). By using an identical task with Gallagher et al. (2000), Vogeley et al. (2001) examined brain activation when participants were answering ToM questions, self-perspective related questions and physical questions after reading similar stories. Results again confirmed the role of mPFC in processing mentalizing questions, however, prompting self-perspective related questions increased the activation in the temporal-parietal junction. In a critical review, Gallagher and Frith (2003) first summarized these early findings of neuroimaging studies on mentalizing and recognized that the medial prefrontal cortices, the superior temporal sulci and the bilateral temporal poles as crucial parts of the mentalizing network. Later, Saxe and Powell (2006) further identified the selective role of right TPJ in processing other's mental content. By asking participants to read stories about other's thoughts, physical appearance or bodily sensations in the scanner, they found that the mPFC area actively participates in all these tasks, however, the TPJ only activated when stories are about other's thoughts. So far there seems to be a wide consensus that the mentalizing network includes the medial prefrontal cortex (mPFC), the bilateral temporal-parietal junction (TPJ), the precuneus and the superior temporal sulcus (STS).

Although these neuropsychological studies provide similar findings on the neural mechanism of mentalizing, more recent results indicate that the mPFC and TPJ may associate with distinct mentalizing processes. After Saxe's (2003) initial explorations on TPJ, the specific role of TPJ in mentalizing was further suggested by studies with patients with brain injuries. Both Samson, Apperly, Kathirgamanathan, & Humphreys (2005) and Apperly, Samson, Chiavarino, & Humphreys (2004) reported that patients suffered from temporal-parietal junction injury encountered deficit in false-belief reasoning, regardless of being tested with video or verbal-based false belief stories. Collectively, Van Overwalle (2009) summarized the neuroimaging results on mentalizing, and argued that mPFC may be involved in more enduring states of mentalizing but TPJ may play a more critical role when the mental states are more transient, e.g. processing other's visual perspectives.

5.1.3 The role of TPJ in visual perspective taking

Until now the neuropsychological evidence for VPT is still quite limited. However, current studies seemed to constantly report the active involvement of TPJ in either inhibiting the egocentric perspective, or facilitating other's perspective. Using Samson's altercentric intrusion task, Schurz et al.(2015) tested in the scanner how multiple brain regions react to conditions where participants experienced congruent or incongruent perspectives with an agent when counting the number of discs. They found activations in the right TPJ, the ventral mPFC and the ventral precuneus. More specifically, when participants were asked to judge the number of discs from the egocentric perspective, these areas activated differently in the perspective-congruent and perspective-incongruent conditions. McCleery, Surtees, Graham, Richards, & Apperly (2011) used ERP and measured brain activities with the Samson's disc-counting task (Samson et al., 2010). They identified a middle-latency temporalparietal component, which showed the longest latency when participants needed to take an incongruent perspective. Also by analysing the electrophysiological signals, Beck, Rossion, & Samson (2018) rapidly presented participants with pictures used in the Samson's task (at a frequency of 2.5Hz). Importantly, pictures showing the agent holding an incongruent perspective were presented at a lower frequency (0.5Hz). The EEG signal demonstrated a clear 2.5Hz stimulation rate response on the spectrum from the typical medial occipital sites. But more critically, there was a 0.5Hz response signal which couples with the frequency of incongruent perspective stimuli originating from the central parietal lobe region.

Martin, Huang, Hunold, & Meinzer (2019) applied high definition transcranial direct current stimulation (HD-tDCS) to investigate the neural mechanism of VPT using both a level-1 (Samson et al., 2010) and a level-2 VPT task. In the level-2 VPT task participants needed to report how an agent in the pictures of traffic scenes would judge the spatial relationship between two targets. They found that in the level-2 but not the level-1 VPT task, anodal stimulation on rTPJ results in shorter reaction time when participants were reporting from the other person's perspective in the incongruent perspective condition. Whereas anodal stimulation on dmPFC caused an increase in RT under incongruency when reports were based on the egocentric perspective on the level-1 VPT task. They argued that such results suggested that the function of rTPJ is to inhibit the egocentric perspective during perspective-taking, while dmPFC allows one to integrate other's visual input into the self-perspective.

On the other hand, there is other evidence against the role of TPJ in perspective-taking. Dumontheil, Küster, Apperly, & Blakemore (2010) investigated VPT neural mechanism using the Director task but failed to find the involvement of TPJ. In their paradigm, participants were instructed to interact with two directors in the pictures presented on a laptop, wherein the task they need to follow the director's instruction to take a correct object from a set of shelves. One director stands on the side with the participant, so they both can see the same thing. But the other director stands at the back of the shelves. Since some of the shelves have an opaque back, objects placed in these shelves were thus blocked from the director on the back and thus creating perspective incongruency. However, those objects can be seen by both the other director and the participant. With this manipulation, they found that the incongruent perspective condition compared with congruent condition led to the recruitment of the superior dmPFC, the middle temporal gyri and the temporal pole but not TPJ. They claimed that TPJ may be more closely related to VPT in cases where participants need to make behavioural predictions or anticipate action consequences.

Other studies investigate the process of visual perspective taking together with other social cognitive factors, such as imitation or goal-directed movements. Jackson, Meltzoff, & Decety (2006) asked people to watch hand or foot movement pictures either from a first or a third perspective, and found that mere observing from the 3rd-person perspective induces more activities in the lingual gyrus (a lateral part of the occipitaltemporal area). Similarly, in David et al.(2006) participants played a ball-tossing game with two avatars in the scanner. They either played from the 1st-person's or the 3rd-person's perspective. Results indicated that playing from a 3rd-person's perspective yielded more activities in regions such as the left

IPL and parietal–temporal–occipital junction, and the right superior parietal lobe. These results provide insights into how processing other's visual perspective is related to processing movements.

Later on, Mazzarella, Hamilton, Trojano, Mastromauro, & Conson (2012) investigated the role of other's actions during perspective-taking. They conducted a study where participants made left-right judgments of the location of an object either based on their egocentric or another person's perspective. They reported dissociated responses from the dmPFC and the inferior frontal gyrus (IFG), with dmPFC more sensitive to the other person's orientation in the altercentric perspective condition and IFG more sensitive in the egocentric perspective condition. The occipitotemporal cortex also showed higher activation when the person is reaching to the object compared with that in the non-reaching condition, but such activation is independent of perspective conditions. Santiesteban, Banissy, Catmur, & Bird (2015) examined the neural mechanism overlap between VPT, imitation and theory-of-mind by using tDCS by exerting anodal stimulation over right TPJ, left TPJ or the occipital cortex. Results showed stimulation on bilateral TPJ reduces the tendency of perspective taking and imitating others, with no influence on theory-of-mind. These results provided further evidence for the critical role of TPJ in perspective-taking.

5.1.4 The role of TPJ in in-group/out-group effects

In Chapter 4, we introduced the dehumanization effect in social cognition, which describes the phenomenon that we do not attribute equal human essence to other people when processing their social information. Such bias also exists in neural activities, as previous studies indicate that the mentalizing network including bilateral TPJ frequently shows different activation during intergroup cognition. A prominent case is in-group favouritism. Volz, Kessler, & von Cramon (2009) studied the neural mechanism of in-group bias by using a minimal group paradigm (MGP). In their study, participants were allocated into a 'blue' or 'yellow' group and were asked to distribute money to in-group or out-group members by making a choice between different allocation matrices. They found that the bias to give in-group members more money recruited four brain regions: the dMPFC, pACC, TPJ, and precuneus. Also, by using a monetary decision-making paradigm, Telzer, Ichien, & Qu (2015) conducted a cross-cultural donation experiment. In each trial, participants viewed a Chinese or Caucasian-American face then decided the amount they would like to donate. Results from this study showed that individuals who have a stronger group identification and the Chinese participants (compared with Caucasian-American participants) demonstrated increased activation in regions such as vlPFC, ACC, TPJ, and dmPFC, when contributing to the out-group relative to in-group persons. Recently, Schiller and colleagues (2019) studied in-group favoritism in sports fans with the EEG instrument. They adopted an Implicit Association Test (IAT), which measures the participants' implicit attitude, and found that a higher IAT score (greater in-group bias) is closely linked with lower levels of theta current density in the right TPJ. These results indicate that during intergroup interactions, the active involvement of TPJ may take part in modulating biased cognition towards ingroup members. Considering the specific role of TPJ in perspective-taking, a possible explanation for such modulation effect could be that individuals spontaneously take other's perspective to reduce cognitive bias.

5.1.5 Current study and hypotheses

To further explore whether TPJ is related to spontaneous VPT and perspective-selection in the social mental rotation task, we adapted both the VR headset and fNIRS optodes layout to allow measurement of neural activities in the TPJ and adjacent areas. We used the same minimal group manipulation as in Chapter 4, where participants completed a dot-counting task and received a role as either an over-estimator or under-estimator. Then they performed the social mental rotation task in VR with ingroup and outgroup agents while brain activity was recorded with fNIRS. We used two slightly different versions of the social mental rotation task in different blocks. The two-agent version of the task had both an in-group agent and an out-group agent present at the same time, just as in Chapter 4. The one-agent version had only a single agent present on each trial; this was used to make it easier to distinguish the effect of the agent's group membership on brain activity.

We hypothesized that stronger TPJ activities would be observed when participants need to take an altercentric perspective in the one-agent condition. Moreover, compared with the out-group agent condition, in the in-group agent condition we should observe stronger TPJ activation. Similarly, in the two-agent condition, we should observe more involvement from TPJ when participants were taking an in-group agent's perspective. As no neuroimaging studies have compared how VPT-related brain areas respond to one-perspective or multi-perspective scenarios, we would leave this as an open question and explore possible evidence.

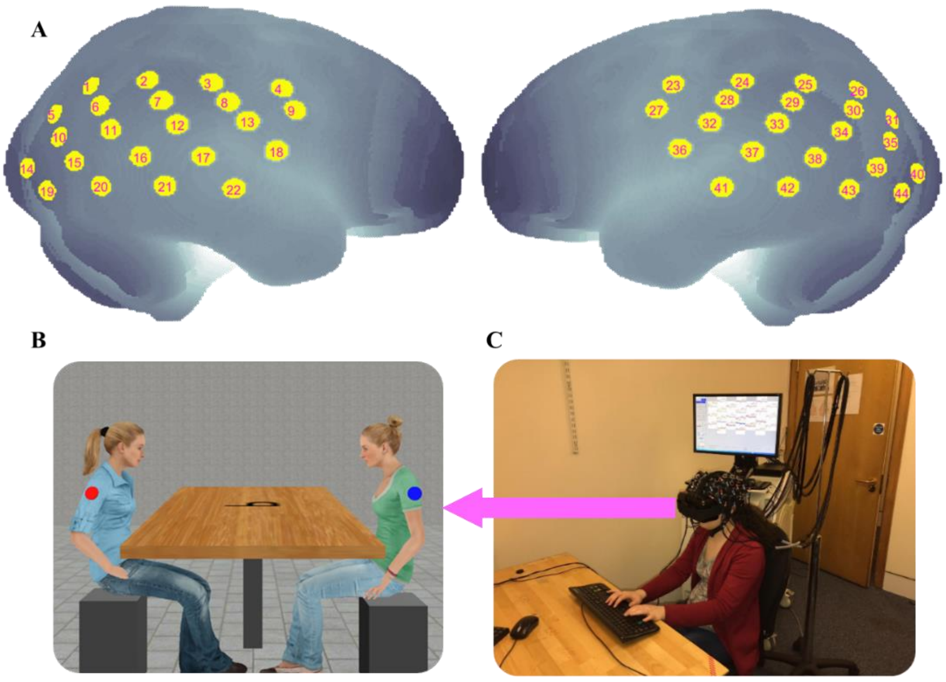


Figure 5-1. Channel layouts of the fNIRS setup. (A). Participants wear Oculus Rift DK2 and an EEG cap wherein 30 fNIRS optodes were plugged in on both hemispheres (C). From the Oculus headset, participants saw the VR setting of the social mental rotation task (B). Figure B displays an example trial from the two agent conditions.

5.3 Method

Participants

32 right-handed participants took part in this experiment (15 females, mean age = 25.1, SD = 6.36). Participants were recruited from two UCL-associated psychology databases, and were required to have the Latin alphabet as the basis of their first language and normal or normal-to-corrected vision. Participants were paid at a rate of £7.5 per hour to compensate for their time. The study is under the ethical approval from the UCL Research Ethics Committee and conformed to the 1964 Declaration of Helsinki.

Materials, VR setup & Experiment design

The dot-counting task was again used for the minimal group manipulation. Participants completed all the dot-counting trials without the VR headset and the fNIRS instrument. In each trial, they needed to estimate the number of white dots on the screen, within the range of 50-250. Then after 10 trials, they were informed of their roles as an over-estimator or under-estimator, then rated to what extent they agree with the results and received a coloured sticker according to their roles.

The VR setting for the social mental rotation task was created by in Vizard 5.0 (Worldviz, USA). Participants wore the Oculus Rift DK2 and saw a virtual room where a wooden table was placed with either one or two female agents sitting on the left or right (Figure 5-1.) Both agents had a European appearance and moved according to the Vizard 'quiet sitting' animation with breathing and small movements; the agent's head was oriented so that they were looking towards the center of the table where the stimuli appeared. The agents wore a coloured sticker on their arms to indicate their roles as an over-estimator or an under-estimator, and their roles were balanced across participants. As in Chapter 4, we used four asymmetric letters as our stimuli. In each trial, one letter randomly selected from the set F, R, P, and G was presented in the center on the table in Times New Roman font.

Each participant completed 24 miniblocks from four different conditions (see Figure 5-2.): named 'In', 'Out', 'BIn', 'BOut'. In the 'In' and 'Out' conditions, only the in-group ('In' condition) or the out-group ('Out') condition agent sat in the room by the table, and their sitting positions are balanced across miniblocks. There were 9 trials in each miniblock, with 8 valid trials and 1 control trial. In half of the eight trials, the letter presented would be oriented towards the agent, and in the other half they were oriented away. In the control trial, the letter would face towards the participants, and we added this trial to increase the uncertainty of the letter orientation.

In the 'BIn' and 'BOut' conditions, both the in-group and the out-group agents were in the scene, with their sitting positions balanced across miniblocks. Each miniblock contains 8 trials. Unlike the manipulation in Chapter 4, in the 'BIn' condition, letters would be oriented towards the in-group agent in six trials, towards the out-group agent in one trial and to the participant in one trial, and vice versa in the 'BOut' condition. We purposely adopted this unbalanced trial organization to enhance block-level differences between the BIn and BOut blocks. If the letter stimuli in the two-agent blocks were oriented towards each agent equally often, it would not be possible to create a block-level analysis of the preference for an ingroup or outgroup agent. By adopting an unbalanced design, we aimed to better separate out brain activities relevant to adopting an ingroup agent's perspective (in the ingroup-advantage condition) or to suppressing an outgroup agent's perspective (in the outgroup-advantage condition), and relate such results with any behavioural differences. Thus, there are more trials in the two-agent conditions where letters were oriented towards an agent. We thus included more miniblocks from the one-agent condition. A detailed illustration of miniblock structure is presented in Figure 5-2.

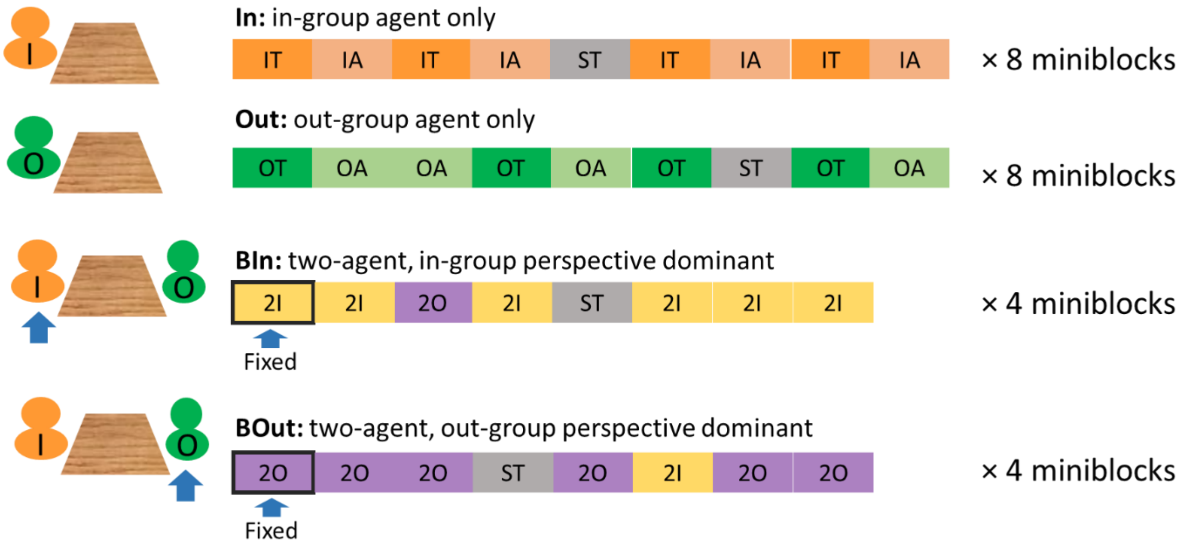

I
O
In: in-group agent only
IT IA IT IA ST IT IA IT IA
× 8 miniblocks
Out: out-group agent only
OT OA OA OT OA OT ST OT OA
× 8 miniblocks
BIn: two-agent, in-group perspective dominant
2I 2I 2O 2I ST 2I 2I 2I
Fixed
× 4 miniblocks
BOut: two-agent, out-group perspective dominant
2O 2O 2O ST 2O 2I 2O 2O
Fixed
× 4 miniblocks

Figure 5-2. The trial design of each miniblock. Meanings of the abbreviations in the figure: **I:** Ingroup **O:** Outgroup **IT:** Ingroup towards **IA:** Ingroup away **ST:** Towards the participant **OT:** Outgroup towards **OA:** Outgroup away **2I:** Towards ingroup in the two-agent block **2O:** Towards outgroup in the two-agent block. In the two-agent conditions, the first trial was always fixed to be the dominant type of trials in this block (marked by the blue arrow).

In the end, there were 8 miniblocks for each one-agent condition, and 4 miniblocks for each two-agent condition. The 24 miniblocks were fully randomised and split into two halves, with a resting period of at least 30 s in-between. We also varied the inter-trial-interval (ITI) to make sure each miniblock lasts for about 32s.

Procedure

Participants gave written consent to take part. Upon their arrival at the lab, participants first received preparation for the fNIRS testing, including measurement of their heads, fitting the fNIRS cap and digitization of the locations of the fNIRS optodes. We made these preparations in advance to guarantee the strongest group effect while participants were doing the social mental rotation task, otherwise participants may forget their group identity.

After the fNIRS cap was fitted, participants first completed ten trials of dot-counting task to be assigned into the over-estimator or under-estimator group. The dot-counting task took about 10 min to finish and then participants were given a coloured sticker according to their role. Participants were then introduced to the two agents (named Lucy and Ellie). Participants were told that the agents were two real participants who completed the same task on a previous day, and one of them was an over-estimator and the other an under-estimator.

Then they were instructed to put on the HMD and get familiar with the room. During the experiment, the two agents wore coloured stickers indicating their group. Both participant's and the agents' labels as an under-estimator/over-estimator were balanced across participants. Agents' sitting positions were balanced across miniblocks.

At the beginning of each block, depending on the condition, either one or two agents sat naturally in the virtual room by the table for 5.5 seconds to let participants adapt to the scene and get familiar with the positions of the agent(s). A red dot then appeared at the center of the table for 500 ms indicating the upcoming letters. On each trial, a letter appeared in the center of the table, orienting towards one agent or towards the participant (on the control trial). Participants were instructed to press key 'J' if the letter was canonical and 'F' if it was mirror-reversed. After the keyhit, the letter disappeared and there was an inter-trial interval of 2.5-3.5 s. The room and agents all remained visible during the inter-stimulus interval, to maintain the feeling of being in a real location with real people.

Each miniblock lasts about 32 s with a break of 15 - 20 s in-between. After 12 miniblocks, participants could take a rest for at least 30 s. After all the computer-based tasks, participants were provided with a questionnaire which aimed to clarify their knowledge and strategies on the task, as well as their attitude towards the two agents in the final part. The whole testing period took about 25 minutes to finish.

fNIRS Brain Measures

We used the Shimazu LabNIRS system (Shimadzu Corp., Kyoto, Japan) to measure brain responses. The fNIRS system was set up with 16 emitters (each emitter contained three light sources at 780 nm, 805 nm and 830 nm wavelengths) and 14 light detectors to obtain measurements at a sampling rate of 37.4 Hz. On each hemisphere, 8 emitter and 7 detector optodes were arranged on a headgear in a 3*5 array with 3 cm spacing to cover a wide patch of the scalp over the temporal-parietal junction, the superior temporal sulcus, the inferior parietal lobule and the anterior part of the occipital lobe (Figure 5-1.) The custom headgear was modified from an adult size large EEG Electro-Cap (Electro-Cap International, Inc., Eaton, OH). Optodes were vertically inserted into rubber grommets within the cap to make sure they stay in a fixed position during the task.

To place the fNIRS cap in a reliable way across all the participants, we followed the procedure introduced in a previous study (Pinti et al., 2015). Specifically, the cap was placed with the center of its middle-front point marked on the participant's head as the 10% of the

Nasion-Inion distance from the Nasion (Pinti et al., 2015). For the participant comfort, the cap was placed as close to the participant's scalp as possible. To make sure the fNIRS signal we recorded is from the target brain areas such as the bilateral TPJ, sources and detectors were arranged based on a 3 × 5 configuration on each side of the head right above the participant's ears. Thus, after placing on the cap, we first obtained the coordinates of the grommets within the targeted region and the anatomical landmarks by using a 3D magnetic digitizer (Patriot, Polhemus, Vermont). We then used the NIRS_SPM package (Ye, Tak, Jang, Jung, & Jang, 2009) to obtain the MNI coordinates for the location of each source and detector. The MNI coordinate of each fNIRS channel was determined by the mid-point location between each pair of a source and a detector, then the anatomical locations of the channels were located according to a standard brain template. The resulting MNI coordinates are listed in Table 5-1.

Table 5-1.MNI coordinate for each channel.

	MNI coordination (mm)			Standard deviation (mm)		
	x	*y*	*z*	*x(sd)*	*y(sd)*	z(sd)
CH1	45	-75	46	4.3	4.3	6.8
CH2	59	-52	50	3.9	5.9	5.1
CH3	65	-26	49	2.4	7.2	4.2
CH4	60	2	46	2.7	8.9	5.5
CH5	33	-89	31	3.8	4.1	8.4
CH6	53	-71	37	3.5	4.3	7.2
CH7	65	-46	40	2.9	6.3	5.7
CH8	68	-20	40	1.3	7.9	4.6
CH9	63	8	35	3.1	9.7	5.8
CH10	43	-88	21	3.7	3.7	7.9
CH11	59	-67	25	3.3	5.1	7.0
CH12	69	-40	28	2.0	6.6	6.2
CH13	69	-12	30	0.9	8.7	5.5
CH14	31	-100	5	3.6	2.9	6.3
CH15	52	-81	9	2.9	4.1	6.9
CH16	67	-55	13	3.0	6.0	6.4
CH17	72	-30	13	1.1	7.9	6.5
CH18	68	0	16	2.3	9.9	6.1
CH19	42	-92	-6	3.9	4.0	6.2
CH20	58	-70	-2	3.8	6.4	6.2
CH21	71	-45	-2	1.8	7.8	5.2
CH22	72	-17	-3	2.2	9.2	5.9
CH23	-56	-2	50	3.7	11.4	6.9
CH24	-60	-30	52	3.6	10.1	5.1
CH25	-55	-55	50	4.4	9.3	5.5
CH26	-43	-76	46	5.6	7.4	7.5
CH27	-59	5	38	3.4	12.7	7.8
CH28	-64	-24	43	2.2	11.3	5.9
CH29	-62	-50	42	3.7	9.9	6.1
CH30	-50	-74	37	5.3	8.3	7.7
CH31	-34	-90	31	6.1	6.4	9.2
CH32	-66	-17	32	1.6	12.5	6.8
CH33	-67	-44	32	2.7	11.1	7.0
CH34	-57	-69	27	5.4	9.7	7.7
CH35	-42	-89	21	6.4	7.0	8.6
CH36	-67	-5	19	3.5	13.3	8.1
CH37	-69	-34	18	1.2	12.0	7.7
CH38	-64	-59	14	3.9	10.0	7.1
CH39	-50	-84	9	6.2	8.7	7.5
CH40	-32	-99	5	6.7	5.2	6.9
CH41	-70	-22	0	3.0	11.7	8.3
CH42	-69	-48	0	2.2	10.9	6.7
CH43	-57	-73	-1	5.8	9.8	6.5
CH44	-41	-93	-4	6.5	6.6	6.4

After the dot-counting task, optodes were then plugged into the grommets according to the 3 × 5 configuration on each side. In total, 16 sources and 14 detectors together created

44 measurement channels. After all the optodes were plugged in the grommets, a black cloth was used to cover the optodes during the signal quality check and the data collection stage to block the room light. Ten-foot optical fibers carried light to and from the head probe. Before the task began, we first conducted initial measurements to check signal strength and quality. Attenuation values of the source-detector pairs less than 60 db or greater and 150 db were automatically adjusted by the system, or by manually adjusted by re-plugging the optode to the grommet, or by displacing any impeding hair. Experimental recording proceeded if all active channels had signal strength between 60 db and 150 db or if the signal channel cannot be improved any more. We estimated the cortical sensitivity of each channel in our probe through simulations of photon migration in a realistic 3D head model using the mesh-based Monte Carlo photon migration simulation algorithm implemented in AtlasViewer version 1.3.9 (implemented in Homer2 v 1.5.2)

fNIRS data Analysis methods

We conducted pre-processing of fNIRS data using the Homer2 toolbox (Huppert, Diamond, Franceschini, & Boas, 2009). The raw intensity data were first converted into changes in optical density using the *hmrintensity2OD* function, then was corrected for motion artefacts based on the wavelet-based method (function in Homer 2: *wavelet_motion_correction*, iqr = 1.5, see Molavi & Dumont, 2012). We then applied a band-pass filter method (function in Homer 2: *hmrBandpassFilt*; order: 3rd; band-pass frequency range [0.01 0.50] Hz) to remove physiological noise such as heart rate, and low-frequency noise slow trends in the data. Then, the modified Beer-lambert law was applied to compute the concentration change of HbO_2 and HbR (function in Homer2: *hmrOD2Conc*; assuming a fixed DPF of 6, see Yücel, Selb, Aasted, Lin, Borsook, Becerra, & Boas, 2016). The pre-processed data was then visually inspected to assess signal quality. Channels with detector saturation or poor optical coupling (marked as a lack of heartbeat signal (frequency at 1Hz -1.5 Hz)) were then excluded from further analysis. To localize the functional activation on the basis of one signal including changes of both HbO_2 and HbR, we then applied the correlation based signal improvement (CBSI) method (Cui, Bray, & Reiss, 2010). This approach has the potential to reduce false positives at the statistical inference stage (Tachtsidis & Scholkmann, 2016). Finally, data were downsampled to 5Hz. After preprocessing was complete, we analysed data using a channel-wise GLM approach. We used two different GLM models, one to fit data at the trial level and a different one to fit data at the block level. We describe each in turn.

Trial-based analyses of the single-agent blocks

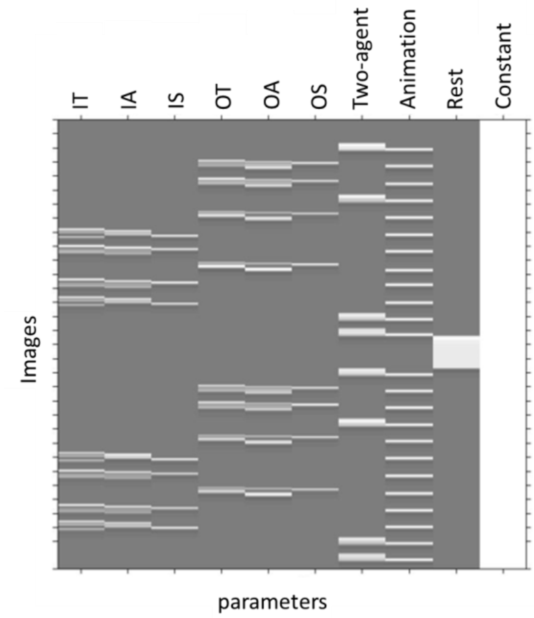


Figure 5-3. The design matrix of the single-agent miniblock analyses. Meanings of the abbreviations in the figure: **IT:** Ingroup towards **IA:** Ingroup away **IS:** Towards the participant in the one-agent ingroup block **OT:** Outgroup towards **OA:** Outgroup away **OS**: Towards the participant in the one-agent outgroup trial **Two-agent:** all trials in the two-agent blocks **Animation**: the 5.5s in the beginning of each miniblock **Rest**: the rest period between the first and second half of the experiment **Constant**: Constant factors

The aim of this analysis was to test if the social mental rotation task activates VPT related brain areas, and to explore ingroup/outgroup differences in perspective taking when only one agent was present. First, we constructed a design matrix for each participant which models the individual trials in the single agent blocks (IT: in-group towards; IA: in-group away; IS: in-group self, OT: out-group towards; OA: out-group away; OS: out-group self). We added additional regressors to fit for all the trials in the two-agent conditions (Two-agent), for the animation period before each miniblock (Animation), for the interval in the middle of the experiment (Rest) and for the constant factors (Constant) (see Figure 5-3. for the design matrix).

All regressors were computed convolving the event epochs corresponding to each trial with the canonical Hemodynamic Response Function (HRF) and were used to fit the fNIRS activation signals. To do this, we applied first-level (or single-subject) channel-wise GLMs (using SPM_fNIRS toolbox https://www.nitrc.org/projects/spm_fnirs/) to the 44 channels' activation signal for each participant, down-sampled to 5Hz. From these, we obtained 10 beta values for each channel of each participant. Next, we created second-level GLMs on these beta-values to test the following contrasts:

Away > Towards > Self contrast to reveal which brain areas in the measured regions linearly increase with the difficulty of mental rotation;

Towards > Away contrast to reveal out the brain areas related to perspective taking in the one agent condition;

Away & Towards > Self contrast to reveal which brain areas are related to mental rotation;

Ingroup: Away –Towards –Self > Outgroup: Away –Towards –Self to reveal if any brain area involves in the interaction effect between Group factor and Direction factor.

The beta-values obtained from these contrasts were then entered the second level GLMs. The second-level GLMs were implement in Matlab (2016a) using the 'fitglme' and 'compare' functions, to obtain the significant p-values following a model comparison approach. This means that we would generate a model of the betas which includes the contrast of interest (eg. Equation 1) and a baseline model without that contrast (eg. Equation 2) and then compare the two models. If the test model gives a better fit to the data, we consider the contrast to be reliable.

$$\text{beta-values} = \beta_1\text{Group} + \beta_2\text{Direction} + \beta_3\text{Group} \times \text{Direction} + \varepsilon \qquad (1)$$

$$\text{beta-values} = \beta_1\text{Group} + \beta_2\text{Direction} + \varepsilon \qquad (2)$$

For each of these contrasts, the model comparison gives returns p-values for each contrast and each channel. To correct for multiple comparisons, based on a previous approach (Southgate et al., 2014), we consider two or more adjacent channels with p-values smaller than 0.05 as significant results.

Miniblock-based analyses

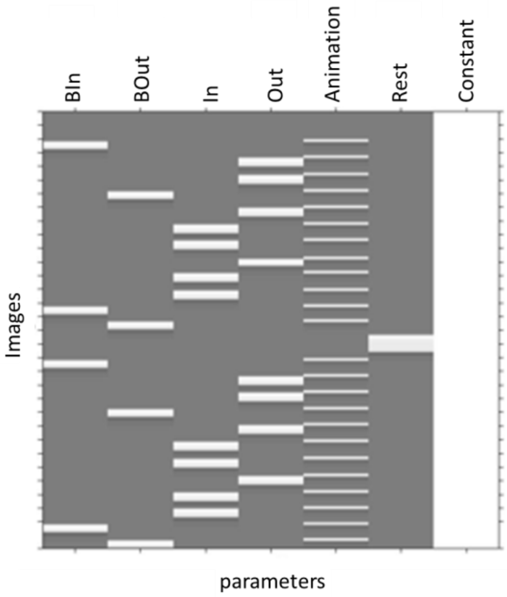


Figure 5-4. The design matrix of the miniblock-based analyses. Meanings of the abbreviations in the figure: **BIn:** Ingroup advantage two-agent miniblocks **BOut:** Outgroup advantage two-agent miniblocks **In:** Ingroup agent only condition **Out:** Outgroup agent only condition **Animation**: the 5.5s in the beginning of each miniblock **Rest**: the rest period between the first and second half of the experiment **Constant**: Constant factors.

These analyses aim to uncover brain activities relevant to the differences of miniblocks. Similar to trial-based analysis, we first constructed design matrices for different types of miniblocks. Seven regressors were modeled for the design matrix, in which four of them are for encoding the different miniblock types (‘In’: in-group, ‘Out’: out-group; ‘BIn’ : two-agent in-group; ‘BOut’ : two-agent out-group), one for the animation period before each block (Animation), one for the interval in the middle of the experiment (Rest) and the last one for the constant factor (Constant) (see Figure 5-4 for the design matrix).

Then beta-values for the seven regressors were first obtained for each participant on each channel by using the SPM_fNIRS toolbox. Following the same method described above, we fit the beta-values with a second-level GLM in using a model comparison approach. We fitted the beta-values for each contrast and each channel into a GLM containing the aimed contrast and compared it with a more simplified model, and then obtained the p-values for each contrast on each channel. The contrasts we are interested in are:

In > Out to reveal which part in the measured brain regions is involved in preferentially taking the in-group agent’s perspective;

BIn > BOut to reveal which part in the measured brain regions is involved in selecting the in-group agent’s perspective in the two-agent setting;

Two-agent > One-agent to reveal which part in the measured brain regions has stronger activation when participants were facing two agents than only facing one agent;

AllIn > AllOut to reveal which part in the measured brain is involved in generally selecting the in-group agent's perspective over the out-group agent's perspective.

For the significant results, similar to the trial-based analyses, we considered the activation patterns where two or more adjacent channels with $p < .05$ as significant results (Southgate, Begus, Lloyd-Fox, di Gangi, & Hamilton, 2014)

5.4 Results

We excluded data both according to the fNIRS signal quality and participants' behavioural results. For all the 32 participants who completed all the tasks, 5 participants were removed from further analyses because of too many artefacts on the fNIRS signal. One participant was excluded because of below chance (and below 3 SDs from the mean) overall accuracy. One participant was removed from all analyses related to the two-agent conditions because of below chance (and below 3 SDs from the mean) accuracy on this task. Thus, there was a valid sample size of 26 for analyses with the one-agent conditions, and 25 for those with the two-agent conditions.

Behavioural results of the social mental rotation task

Data collected on participants' behavioural performance were separated into one-agent and two-agent conditions for analyses.

One-agent conditions

A 2×2 repeated-measurement ANOVA was applied to analyze accuracy for all trials and reaction time (RT) for correct trials after excluding extreme values (± 3 SDs), with Letter-direction (towards; away), and Agent (in-group, out-group) as within-subject factors. For accuracy, there is a significant interaction (F (1, 25) = 5.78, p = .024, $\eta_p{}^2$ = 0.19), no main effects is significant (for Agent: F (1, 25) = 2.90, p = .101; for Letter-direction: F (1, 25) = 1.01, p = .325). Post-hoc analyses showed that in the in-group condition, participants performed significantly better when the letter faced towards the agent (t (25) = 2.079, p = .048), but such difference is not significant in the out-group agent condition (t (25) = -1.046, p = .305) (Figure 5-5.A).

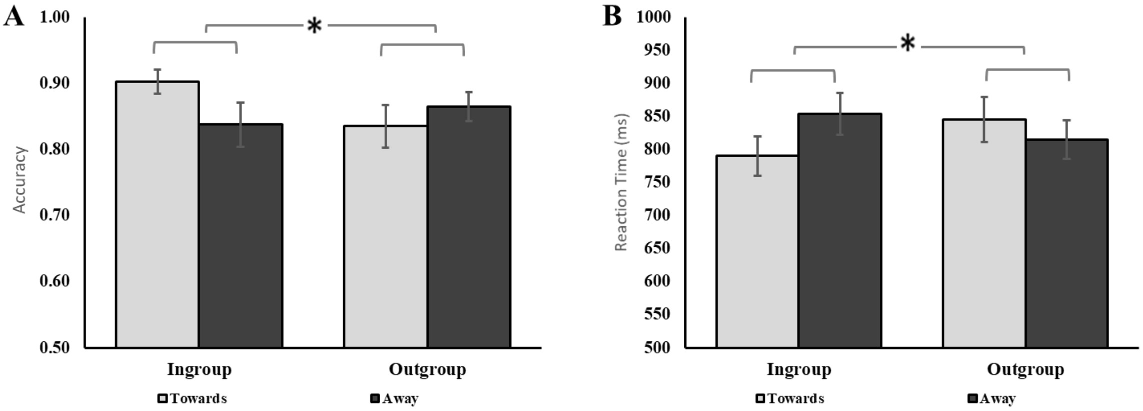


Figure 5-5. Results of the one-agent miniblock analyses. (A) Results on accuracy of recognising letters F, R, P and G; (B) Results on reaction time (ms) of recognising letters F, R, P and G.

For RT, a significant interaction was found for Letter-direction (towards; away) × Agent (in-group, out-group) (*F* (1, 25) = 14.83, *p* = .001, $\eta_p{}^2$ = 0.37), when letters faced towards the in-group agent they were identified faster by the participant than when they faced away (*t* (25) = -3.774, *p* = .001). But such difference disappeared when letters faced towards the out-group agent (*t* (25) = 1.713, *p* = .099). There was no significant main effect for either of the two factors (Figure 5-5.B)

Two-agent conditions

As in Exp.1 of Chapter 4, for the two-agent conditions, we included Agent and Letter-type as the two factors in our analyses. A 2×2 repeated-measurement ANOVA was applied to analyze accuracy for all trials and reaction time (RT) for correct trials after excluding extreme values (± 3 SDs), with both Agent (in-group, out-group) and Letter-type (normal, mirror-reversed) as within-subject factors.

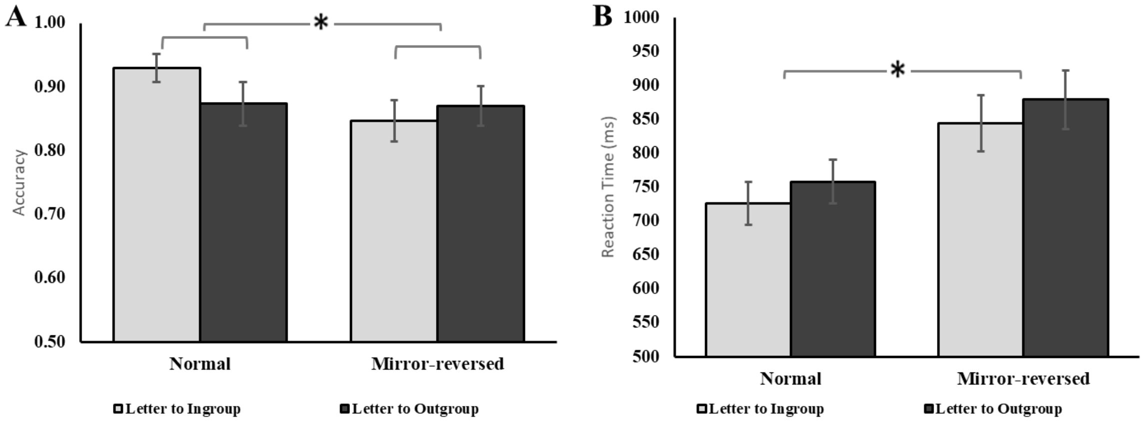


Figure 5-6. Results of the two-agents condition. (A) Results on accuracy of recognising letters F, R, P and G; (B) Results on reaction time (ms) of recognising letters F, R, P and G.

For accuracy, results revealed a significant interaction between Agent* Letter-type (*F* (1, 24) = 4.90, *p* =. 037, η_p^2 = .17). There was no main effect for Agent (*F* (1, 24) = 0.47, *p* = .499), or for Letter-type (*F* (1, 24) =3.17, *p* = .088). Post-hoc analyses showed that in the in-group condition, participants performed significantly better when the letter faced towards the agent (*t* (24) = 3.397, *p* = .002), but such difference is not significant in the out-group agent condition (*t* (25) = 0.095, *p* = .925) (Figure 5-6.A)

For RT, no significant interaction was found between Agent * Letter-type (F (1, 24) = 0.01, p =. 924), also there was no main effect for Agent (*F* (1, 24) = 2.16, *p* = .155). Only a significant main effect for Letter-type (*F* (1, 24) =27.47, *p* < .001, η_p^2 = .53) was found, participants recognized normal letters faster than mirrored letters (Figure 5-6.B).

fNIRS results

Trial-based analyses

For contrasts based on different trial types in the single-agent miniblocks, several groups of channels showed significant results. For the *Away* > *Towards* > *Self* contrast , the left inferior parietal lobe (IPL, CH28) and the anterior part of the left temporal-parietal junction (TPJ, CH33) were more activated when there was an increased demand for mental rotation. But when collapsing the Towards trials with the Away trials and compared then with the Self trials (*Towards trials & Away trials > Self trials*), in addition to the activation in the same areas on the left hemisphere, there were also increased activities on the right premotor cortex (CH4), the right IPL (CH8), and an anterior part of the right TPJ. This result suggested that these areas may play critical roles in cognitive tasks which involve mental rotation. The *Towards* > *Away* contrast involves deactivation of the superior parietal lobule (CH2, 3), a posterior part of TPJ (CH6) and angular gyrus (CH10) on the right hemisphere, indicating these regions may take parts in when participants viewing items from an agent’s perspective. Lastly, when considering the Group * Letter-Direction interaction (*IA - IT – IS > OA - OT – OS*), a large group of channels (CH39, CH40, CH43, CH44) covering the left occipital-temporal junction showed increased activity (Figure.5-7).

Miniblock-based analyses

To our surprise, the contrasts on the miniblock level didn’t reveal any significant channel groups. Only a few single channels showed p values under 0.05. However, as we consider two or more adjacent channels show a significant level of .05 of uncorrected p-value as meaningful results (Southgate et al., 2014), no area in the measured brain regions is closely related to these contrasts.

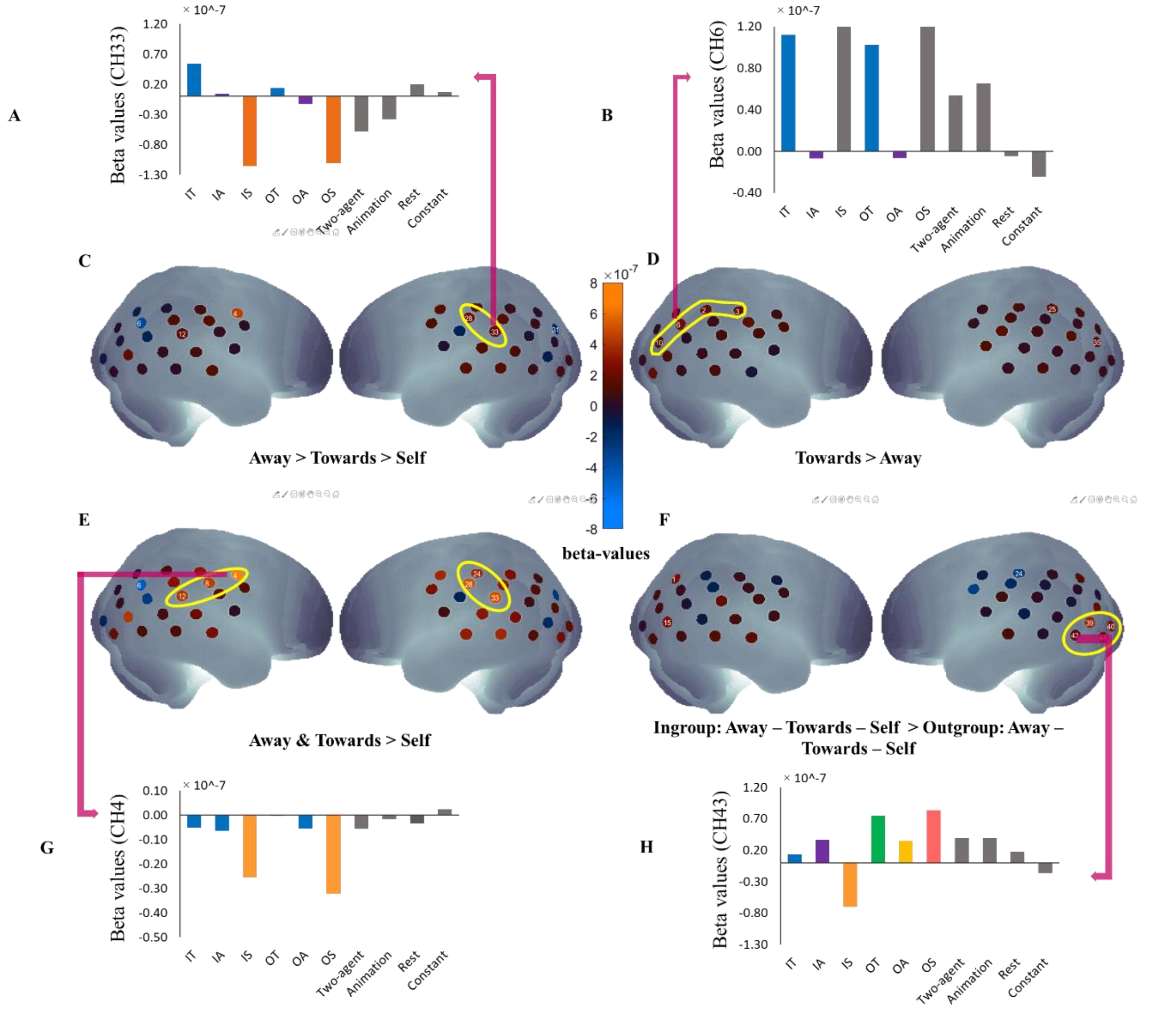


Figure 5-7. The fNIRS results of the one-agent miniblock analyses. Meanings of the abbreviations in the figure: IT: Ingroup towards IA: Ingroup away IS: Towards the participant in the one-agent ingroup block OT: Outgroup towards OA: Outgroup away OS: Towards the participant in the one-agent outgroup trial Two-agent: all trials in the two-agent blocks Animation: the 5.5s in the beginning of each miniblock Rest: the rest period between the first and second half of the experiment Constant: Constant factors.

5.5 Discussion

In the current study, we investigated the neural activity patterns related to the social mental rotation task we implemented in Chapter 4 by using fNIRS together with VR. The behavioural results replicated our previous findings in Chapter 4, by showing letters (normal letters in the two-agent conditions) oriented towards an in-group agent were recognized more accurately. We believe the consistent pattern among this current finding, the results from the pilot study and the first study in Chapter 4 could support that the two-agent perspective taking task is an effective measurement to test participants' bias towards in-group members while taking other's perspective.

In terms of the neuroimaging results, for contrasts within the one-agent conditions, we identified several regions involved either in the process of mental rotation, or perspective taking, or the in-group bias of perspective-taking. But for the contrasts related to the two-agent conditions, and the comparison between the one-agent and two-agent conditions, we failed to find out any relevant brain regions. In the discussion, we will illustrate the implications of these current findings, the possible reason for failing to reveal brain areas relevant to the two-agent conditions, the potential of implementing fNIRS in social psychological studies and the limitations of the current study.

5.5.1 Brain areas related to perspective taking and mental rotation in the current task

Our data show some interesting findings in brain regions related to perspective taking in the one-agent conditions. The behavioural results showed no main effects for Letter-direction (towards; away) on either accuracy or RT, indicating that partiticpants may have invested equivalent effort to the Towards and Away trials. However, the fNIRS results indicate this might not be the case. Results show that the right superior parietal cortex, the right temporal-parietal junction and the occipital-temporal-parietal junction were more activated when participants viewed letters from the agent's perspective (Figure 5-7). Thus, although the behavioural performance for these two conditions may be the same, participants invested different mental effort in these two different types of trials. When rotating the self to a person's position, more brain regions are involved.

Such result patterns are in line with findings in previous studies, which reported increased rTPJ activities in the perspective taking conditions. But it is worth noting that the contrast we conducted here is not fully alike with that in previous studies. In previous tasks, researchers usually compared the perspective taking condition with the egocentric condition, and in most cases they put the egocentric and altercentric perspective in competing roles in the experimental task. For example, in Samson's disc-counting task, the numbers of discs that the participants can see and that the agent can see are both interpretable in such scenarios. As in the number verification task (Elekes et al., 2017; A. Surtees, Samson, et al., 2016), both the visual content of the egocentric perspective (eg. number '6') and the altercentric perspective (eg. number '9') are meaningful. Thus, in these tasks, participants usually needed to inhibit their responding from the egocentric perspective when taking an altercentric perspective. In our task, however, the stimuli are less ambiguous can only be interpreted from one perspective. This means that participants do not have to solve the perspective conflict between the 'self' and 'other', instead they can always benefit from taking the agent's perspective on the 'Towards' trials. Similarly, on the 'Away' trials, an egocentric perspective does not help with identifying the letters. Thus, the major difference between our task and previous ones is the lack of inhibiting the egocentric perspective. Given the presence of this difference, results consistently show activation in the rTPJ when there is an increasing need to adopt another's perspective. Such results thus further confirmed that the rTPJ is a critical brain region closely related to perspective-taking.

As a mental rotation task, our current task also reliably activates several brain regions frequently reported in previous mental rotation studies, such as the intraparietal sulcus and the medial superior precentral cortex (for a review see Zacks, 2008). Compared with trials where there is no need for mental rotation (the *Self* trials), the bilateral superior and inferior parietal cortices showed more activation when participants were doing the *Towards* and *Away* trials. These results are consistent with findings that the superior part of the parietal lobule is more sensitive to the difference between various mental rotation conditions (Gogos et al., 2010). The linear comparison (*Away* > *Towards* > *Self*) of mental effort for switching to the agent's perspective or to an empty location also showed increased activation in both the left superior and inferior parietal cortex. As these areas were previously found to relate to the analog representations and motor simulation process of mental rotation, these results may suggest that participants used the strategy to rotate their frame-of-reference when doing the task.

5.5.2 The bias towards taking an ingroup agent's perspective

Another interesting finding, although a bit unexpected, was that the Group × LetterDirection interaction activated a region in the extrastriate visual cortext (EVC) in the one-agent condition. When an ingroup agent was present, letters facing towards or away from the agent caused stronger activation in this area than that in the 'self' trials, but such difference disappeared when an outgroup agent was present. Although the EVC mainly processes visual information, previous studies have shown that many substrates in this area, including the fusiform body area (FBA) and extrastraite body area (EBA), play a crucial role in social cognition and are closely related to intergroup cognition.

Compared with other areas in the visual cortex, the FBA and EBA become significantly activated when processing body-related information, such as static figures, photos or silhouettes of humans, etc (Amoruso, Couto, & Ibáñez, 2011). The function of FBA slightly differs from that of EBA in that it’s more dedicated to process holistic information. But the main characteristics of FBA and EBA are their important roles in body identity processing (Urgesi, Candidi, Ionta, & Aglioti, 2007), action perception and understanding (Katsumi & Dolcos, 2018), embodied cognition and self-other distinction (Arzy, Thut, Mohr, Michel, & Blanke, 2006), etc. Recently, research showed that the EBA may also contribute to integrating social traits with body information, and suggested this feature may act as one of the mechanisms of intergroup cognitive bias (Aleong & Paus, 2010; Greven & Ramsey, 2017; Marini, Banaji, & Pascual-Leone, 2018).

Azevedo and colleagues (2013) compared brain responses when people were empathizing with ingroup or outgroup individuals. They showed black and white participants pictures of a black or white male hand suffering a harmful treatment (deeply penetrated) or receiving a gentle touch, and found that no matter what type of stimulus was given, the EBA has a stronger activation for the same-color hands. The researchers interpreted that the results suggested the EBA may participate in increasing attention towards same-race bodies. Another study investigated how the FBA integrates social traits with ingroup/outgroup bodies. Greven & Ramsey (2017) also used both racial and minimal group manipulation, and asked participants to learn the associations between different body figures and statements of positive, neutral and negative traits. They found that when positive traits were associated with ingroup bodies, as well as when negative statements were with outgroup bodies, the coupling between the right FBA and the left TPJ was stronger than that in other conditions. But other studies presenting the ingroup and outgroup agents simultaneously to the participants failed to discover different activity patterns in either the FBA or EBA. In one study, Katsumi & Dolcos (2018) showed participants videos with virtual black or white agents approaching/avoiding, handshaking/no-handshaking with each other. Results showed enhanced activity of the EBA in all social encountering conditions than in the control condition, however there was no difference when the encounter happened in the same-race or different-race. Similarly, Azevedo et al.(2013) presented soccer fans short video clips when two soccer teams were playing with each other, and analyzed the intersubject correlation (ISC) when fans of different teams were watching these video segments. They found that the EBA in overall showed stronger activities when participants were watching the game video, compared with the baseline condition, however the ISC did not differ between within or between group analyses. These results suggest that the EBA and FBA may be more sensitive to contexts where only one ingroup or outgroup agent is present.

In the current study, we extended these findings and showed that the extrastriate visual cortex (EVC) is also involved in the perspective-taking process in the intergroup context. Specifically, in the one-agent condition, when the ingroup agent was present, the EVC is more inhibited when an item is facing toward the 'self', but this difference disappeared in the outgroup condition. Currently we are unable to explain these results based on previous findings, as previous studies either only compared the EBA activation to an ingroup versus an outgroup agent, or to a general agent versus a 'self' figure. No study so far has involved an ingroup, an outgroup and a self agent together. But we hypothesize that as the EBA is usually more activated in the embodiment process, the current results may reflect that when interacting with an ingroup agent, participants had a stronger propensity to take the other's perspective and thus were more disembodied when processing their egocentric perspective. That is, participants were more inhibiting their feelings of their own body and position when interacting with an ingroup agent. Such that when a letter facing towards themselves appeared, they might still be imagining seeing from the ingroup agent's perspective.

It is also worthy to note that the activation of EBA is equivalent in the one-agent and two-agent conditions, suggesting the EBA might not be modulated by the number of human figures to be processed. As previous research usually involved one agent at one time, or didn't compare the condition which has multiple human figures with that only one agent is present, it is hard for us to explain the current result. It is likely that the activation of EBA is related to the attentional focus, as suggested by Grevens and Ramsey (2017), where if the task requires participants to attend to more human bodies simultaneously, enhanced EBA activity might be observed.

5.5.3 Neural activities during perspective selection in the two-agent conditions

Although comparisons within the one-agent conditions showed the social mental rotation task is related to previously reported VPT related areas, these areas failed to show significant differences for contrasts involving the two-agent factors. We consider these null results may actually have two reasons: the difference of brain signal in the two-agent conditions was too subtle or the lack of measurement on the prefrontal cortex.

Unlike in the previous versions in Chapter 4, for the two-agent conditions, we no longer asked participants to switch perspectives-taking between the in-group and the out-group member, but rather arranged the majority of trials (seven out of nine trials in each miniblock) within each miniblock facing towards one specific perspective (resulting in the in-group dominant or out-group dominant conditions). The purpose of this design was to maximum the group effect, since in the two-agent conditions, both agents are present at the same time with identical appearance, thus the group difference might be quite trivial if the letters orient alternatively between the two, and it would be confusing for us to explain any difference in brain activities. However, although we adopted this measurement, it is possible that the brain signal difference between the BIn and BOut conditions may still be very subtle as the two agents co-presented at the same time. Participants might indeed be more prepared to take the ingroup agent's perspective no matter in which condition, thus although it was an outgroup-advantage condition, participants were still more involved in the ingroup agent's perspective, as they did in the ingroup-advantage condition. If this was the case, we thus wouldn't be able to reveal any difference from these two conditions. Such a hypothesis can be supported by interpreting the behavioural and the neuroimaging results together, as although we didn't find any significant results on brain activities in the two-agent conditions, behaviourally participants performed worse when the letters were oriented towards the outgroup agent, suggesting they might still be used to answer from the ingroup agent's perspective when the majority of letters oriented towards the outgroup agent.

The second possible reason for the null result may be the lack of measurement on the medial prefrontal cortex (mPFC). In the introduction, we illustrated that the medial prefrontal cortex plays an important role in inferring other's intentions. Moreover, previous studies also showed that this region is closely related with the social or non-social decision making process and is sensitive to decisional errors. Recent work further showed that the mPFC is actively involved in participants' response selection processes. Critically, the mPFC also has a significant role in modulating group effect (Telzer et al., 2015; Volz et al., 2009). This evidence converges to the stream that mPFC may play a critical modulating role in our current two-agent perspective selection task. However, current VR headsets do not allow us to measure brain activities from the prefrontal cortex. With smaller and more portable VR headsets in the future, it would then be extremely valuable to record activities from these areas when participants are immersed in various contexts and will be interesting to compare brain activities in conditions with or without immersed environments.

5.5.4 Using fNIRS in social psychological studies

In contrast to fMRI and EEG, the most obvious feature of fNIRS is its tolerance to human movements, which allows us to measure participants' brain activities while they are in a VR envioronment. Some VR headsets are suitable to use in the fMRI scanner, however, as participants can't move their heads, they lose the immersive feeling created by VR. Meanwhile, compared with other neuroimaging methodologies such as EEG/ERP or fMRI, fNIRS provides satisfied spatial and temporal resolution and has a high tolerance to movements. These features allow researchers to record brain activities while participants are talking or moving, or record neural activities from young participants such as infants, meanwhile guaranteed the quality of the recorded signal. In that, fNIRS has become more and more popular in social psychological studies which pursuits a natural setting of experiments.

Beyond these points, our current research further showed that fNIRS has the potential to be combined with other novel testing methods such as VR. As VR provides 3D environment information, with portable fNIRS, future studies will be thus able to explore the neural mechanisms while participants are doing spatial navigation, or exploring more natural social scenarios such as having an interview or being in a party. The wearable and portable version of fNIRS also allows for measuring brain-to-brain coupling while participants are interacting with each other, talking or singing together (Liu et al., 2016; Osaka, Minamoto, Yaoi, Azuma, & Osaka, 2014) or even doing group work (Nozawa, Sasaki, Sakaki, Yokoyama, & Kawashima, 2016) thus opens a brand new direction for investigating the nature of human social interaction.

5.5.5 Limitations of the current study

Due to the placement of the VR and fNIRS on the participants head, we were not able to measure brain activities from the prefrontal cortex, which may be the reason lead to the null results in the two-agent conditions. Future studies may consider including measurements from both prefrontal cortex and more posterior regions such as TPJ, STS, IPL or temporal pole to examine if areas such as mPFC are involved in selecting the perspective to take in multi-perspective scenarios. Moreover, our results also suggest that participants might use flexible strategies when spontaneously taking other's perspectives, thus to investigate the mechanism underlying perspective-selection, it may be crucial to keep the uncertainty feature of the direction of upcoming stimuli, thus it is less likely for participants to fix their perspective on one agent.

It is also valuable to develop more ecological paradigms to examine the perspective-selection problem, such as group coordination scenarios or more conflicting situations between different perspectives. Future studies may also consider to test participants' choice under different cognitive load, to reveal if such selection involves domain general process.

5.7 Conclusion

In this study, we investigated brain activities when participants were completing the one-agent and the two-agent social mental rotation task in VR. Critically, we adopted fNIRS measurements to record participants' brain signals in a relatively comfortable setting. Our behavioural data again showed that people have a stronger propensity to process visual information from an in-group member's perspective. The fNIRS analysis related to the one-agent settings revealed several brain regions playing a critical role in VPT in this current task, including the right TPJ, the right inferiorparietal cortex and the superior temporal sulci. Moreover, our results suggested that the activation of the rTPJ in previous VPT tasks may due to a need to inhibit egocentric perspective, and it may be less involved when individuals are taking a social perspective than a non-social perspective. Surprisingly, the comparisons related to the two-agent conditions didn't give us any significant results. We believe future studies with a larger coverage of brain areas including the mPFC may add to explaining the different behavioural results. As the first study combining VR with fNIRS, we believe these results not only providing new understandings on the mechanism of humans' VPT process, but offer new ways to investigate social cognitive problems, and insights into how to implementing new techniques in psychological research.

References

Aleong, R., & Paus, T. (2010). Neural correlates of human body perception. *Journal of Cognitive Neuroscience*, *22*(3), 482–495. https://doi.org/10.1162/jocn.2009.21211

Amoruso, L., Couto, B., & Ibáñez, A. (2011). Beyond extrastriate body area (EBA) and fusiform body area (FBA): Context integration in the meaning of actions. *Frontiers in Human Neuroscience*, *5*(November), 1–3. https://doi.org/10.3389/fnhum.2011.00124

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience*, *16*(10), 1773–1784. https://doi.org/10.1162/0898929042947928

Arzy, S., Thut, G., Mohr, C., Michel, C. M., & Blanke, O. (2006). Neural basis of embodiment: Distinct contributions of temporoparietal junction and extrastriate body area. *Journal of Neuroscience*, *26*(31), 8074–8081. https://doi.org/10.1523/JNEUROSCI.0745-06.2006

Azevedo, R. T., Macaluso, E., Avenanti, A., Santangelo, V., Cazzato, V., & Aglioti, S. M. (2013). Their pain is not our pain: Brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Human Brain Mapping*, *34*(12), 3168–3181. https://doi.org/10.1002/hbm.22133

Beck, A. A., Rossion, B., & Samson, D. (2018). An objective neural signature of rapid perspective taking. *Social Cognitive and Affective Neuroscience*, *13*(1), 72–79. https://doi.org/10.1093/scan/nsx135

Cui, X., Bray, S., & Reiss, A. L. (2010). Functional Near Infrared Spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *Neuroimage*, *49*(4). https://doi.org/10.1016/j.neuroimage.2009.11.050.Functional

David, N., Bewernick, B. H., Cohen, M. X., Newen, A., Lux, S., Fink, G. R., … Vogeley, K. (2006). Neural representations of self versus other: Visual-spatial perspective taking and agency in a virtual ball-tossing game. *Journal of Cognitive Neuroscience*, *18*(6), 898–910. https://doi.org/10.1162/jocn.2006.18.6.898

Dumontheil, I., Küster, O., Apperly, I. A., & Blakemore, S. J. (2010). Taking perspective into account in a communicative task. *NeuroImage*, *52*(4), 1574–1583. https://doi.org/10.1016/j.neuroimage.2010.05.056

Elekes, F., Varga, M., & Király, I. (2017). Level-2 perspectives computed quickly and spontaneously: Evidence from eight- to 9.5-year-old children. *British Journal of Developmental Psychology, 35*(4), 609–622. https://doi.org/10.1111/bjdp.12201

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109–128. https://doi.org/10.1016/0010-0277(95)00692-R

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind'in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11-21. https://doi.org/10.1016/S0028-3932(99)00053-6

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind.' *Trends in Cognitive Sciences*, *7*(2), 77–83. https://doi.org/10.1016/S1364-6613(02)00025-6

Gogos, A., Gavrilescu, M., Davison, S., Searle, K., Adams, J., Rossell, S. L., … Egan, G. F. (2010). Greater superior than inferior parietal lobule activation with increasing rotation angle during mental rotation: An fMRI study. *Neuropsychologia*, *48*(2), 529–535. https://doi.org/10.1016/j.neuropsychologia.2009.10.013

Greven, I. M., & Ramsey, R. (2017). Neural network integration during the perception of in-group and out-group members. *Neuropsychologia*, *106*(May), 225–235. https://doi.org/10.1016/j.neuropsychologia.2017.09.036

Huppert, T., Diamond, S., Franceschini, M., & Boas, D. (2009). Huppert-2009.pdf. *Applied Optics*, *48*(10), 280–298. https://doi.org/10.1016/j.drugalcdep.2008.02.002.A

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2006). Neural circuits involved in imitation and perspective-taking. *NeuroImage*, *31*(1), 429–439. https://doi.org/10.1016/j.neuroimage.2005.11.026

Katsumi, Y., & Dolcos, S. (2018). Neural correlates of racial ingroup bias in observing computer-animated social encounters. *Frontiers in Human Neuroscience*, *11*(January), 1–17. https://doi.org/10.3389/fnhum.2017.00632

Liu, N., Mok, C., Witt, E. E., Pradhan, A. H., Chen, J. E., & Reiss, A. L. (2016). Nirs-based hyperscanning reveals inter-brain neural synchronization during cooperative jenga game with face-to-face communication. *Frontiers in Human Neuroscience*, *10*(MAR2016), 1–11. https://doi.org/10.3389/fnhum.2016.00082

Marini, M., Banaji, M. R., & Pascual-Leone, A. (2018). Studying Implicit Social Cognition with Noninvasive Brain Stimulation. *Trends in Cognitive Sciences*, *22*(11), 1050–1066. https://doi.org/10.1016/j.tics.2018.07.014

Martin, A. K., Huang, J., Hunold, A., & Meinzer, M. (2019). Dissociable Roles Within the Social Brain for Self–Other Processing: A HD-tDCS Study. *Cerebral Cortex*, *29*(8), 3642–3654. https://doi.org/10.1093/cercor/bhy238

Mazzarella, E., Hamilton, A., Trojano, L., Mastromauro, B., & Conson, M. (2012). Observation of another’s action but not eye gaze triggers allocentric visual perspective. *Quarterly Journal of Experimental Psychology*, *65*(12), 2447–2460. https://doi.org/10.1080/17470218.2012.697905

McCleery, J. P., Surtees, A. D. R., Graham, K. A., Richards, J. E., & Apperly, I. A. (2011). The neural and cognitive time course of theory of mind. *Journal of Neuroscience, 31*(36), 12849–12854. https://doi.org/10.1523/JNEUROSCI.1392-11.2011

Molavi, B., & Dumont, G. A. (2012). *Wavelet-based motion artifact removal for functional near-infrared spectroscopy Wavelet-based motion artifact removal for functional near-infrared spectroscopy*. https://doi.org/10.1088/0967-3334/33/2/259

Nozawa, T., Sasaki, Y., Sakaki, K., Yokoyama, R., & Kawashima, R. (2016). Interpersonal frontopolar neural synchronization in group communication: An exploration toward fNIRS hyperscanning of natural interactions. *NeuroImage*, *133*, 484–497. https://doi.org/10.1016/j.neuroimage.2016.03.059

Osaka, N., Minamoto, T., Yaoi, K., Azuma, M., & Osaka, M. (2014). Neural Synchronization During Cooperated Humming: A Hyperscanning Study Using fNIRS. *Procedia - Social and Behavioral Sciences*, *126*, 241–243. https://doi.org/10.1016/j.sbspro.2014.02.395

Pinti, Paol, Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2018). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1–25. https://doi.org/10.1111/nyas.13948

Pinti, Paola, Aichelburg, C., Lind, F., Power, S., Swingler, E., Merla, A., … Tachtsidis, I. (2015). Using fiberless, wearable fnirs to monitor brain activity in real-world cognitive tasks. *Journal of Visualized Experiments*, *2015*(106), 1–13. https://doi.org/10.3791/53336

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & 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. https://doi.org/10.1037/a0018729

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain*, *128*(5), 1102–1111. https://doi.org/10.1093/brain/awh464

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2015). Functional lateralization of temporoparietal junction - imitation inhibition, visual perspective-taking and theory of mind. *European Journal of Neuroscience*, *42*(8), 2527–2533. https://doi.org/10.1111/ejn.13036

Saxe, R., & Powell, L. J. (2006). It's the Thought That Counts. *Psychological Science*, *17*(8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Schurz, M., Kronbichler, M., Weissengruber, S., Surtees, A., Samson, D., & Perner, J. (2015). Clarifying the role of theory of mind areas during visual perspective taking: Issues of spontaneity and domain-specificity. *NeuroImage*, *117*, 386–396. https://doi.org/10.1016/j.neuroimage.2015.04.031

Southgate, V., Begus, K., Lloyd-Fox, S., di Gangi, V., & Hamilton, A. (2014). Goal representation in the infant brain. *NeuroImage*, *85*, 294–301. https://doi.org/10.1016/j.neuroimage.2013.08.043

Surtees, A., Samson, D., & Apperly, I. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, *148*, 97–105. https://doi.org/10.1016/j.cognition.2015.12.010

Tachtsidis, I., & Scholkmann, F. (2016). False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward. *Neurophotonics*, *3*(3), 031405. https://doi.org/10.1117/1.nph.3.3.031405

Telzer, E. H., Ichien, N., & Qu, Y. (2015). The ties that bind: Group membership shapes the neural correlates of in-group favoritism. *NeuroImage*, *115*, 42–51. https://doi.org/10.1016/j.neuroimage.2015.04.035

Urgesi, C., Candidi, M., Ionta, S., & Aglioti, S. M. (2007). Representation of body identity and body actions in extrastriate body area and ventral premotor cortex. *Nature Neuroscience*, *10*(1), 30–31. https://doi.org/10.1038/nn1815

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., … Zilles, K. (2001). Mind Reading: Neural Mechanisms of Theory of Mind and Self-Perspective. *NeuroImage*, *14*, 170–181. https://doi.org/10.1006/nimg.2001.0789

Volz, K. G., Kessler, T., & von Cramon, D. Y. (2009). In-group as part of the self: In-group favoritism is mediated by medial prefrontal cortex activation. *Social Neuroscience*, *4*(3), 244–260. https://doi.org/10.1080/17470910802553565

Ward, E., Ganis, G., & Bach, P. (2019). Spontaneous Vicarious Perception of the Content of Another’s Visual Perspective. *Current Biology*, *29*(5), 874-880.e4. https://doi.org/10.1016/j.cub.2019.01.046

Ye, J. C., Tak, S., Jang, K. E., Jung, J., & Jang, J. (2009). NIRS-SPM: Statistical parametric mapping for near-infrared spectroscopy. *NeuroImage*, *44*(2), 428–447. https://doi.org/10.1016/j.neuroimage.2008.08.036

Yücel, M. A., Selb, J., Aasted, C. M., Lin, P. Y., Borsook, D., Becerra, L., & Boas, D. A. (2016). *Mayer waves reduce the accuracy of estimated hemodynamic response functions in functional near-infrared spectroscopy*. *7*(8), 3078–3088. Retrieved from https://www.osapublishing.org/boe/abstract.cfm?uri=boe-7-8-3078

Zacks, J. M. (2008). Neuroimaging studies of mental rotation: A meta-analysis and review. *Journal of Cognitive Neuroscience*, *20*(1), 1–19.https://doi.org/10.1162/jocn.2008.20013

Chapter 6. General Discussion

In this thesis, through four studies, we investigated the question of how social interactions among the self-agent-object triplet influence people's propensity to engage in mentalizing. We first reviewed the two-system model and evidence supporting or opposing it. Then based on the current need to explain mentalizing issues in real life, a mentalizing triangle model was put forward, which focused on the self, agent and object elements in a mentalizing process, and proposed the 'self-object', 'self-agent' and 'agent-object' relationships as gateways towards spontaneous mentalizing. We examined these hypotheses in two critical mentalizing processes: theory-of-mind (ToM) and visual perspective taking (VPT). ToM refers to the ability that people used infer other's beliefs, intentions and preferences etc. VPT is the ability which enables us to work out other's visual input, e.g. whether they can see something as we do and how an object appears to them. In Chapter 2 & 3, we tested the 'agent-object' hypothesis in ToM, by using an adapted false belief task. In Chapter 4 & 5, the 'self-agent' hypothesis was examined in VPT, which revealed how people select a perspective to take when facing multiple agents. In the discussion, we shall first provide a short summary of the findings of each chapter, then discuss both the theoretical and methodological implications of these results.

6.1 A summary of the experimental chapters

In Chapter 2, as a first step we intended to test whether spontaneous ToM exists when no social interaction is provided. To achieve this goal, we created a minimal social context, where the object is a neutral target and the other person is a mere observer. As reaction time can easily be influenced by a number of non-social factors, we introduced a signal detection measurement to better reveal changes in participants' perceptual process, by calculating two new indices in addition to reaction time and accuracy: the decision-making criteria and the perceptual sensitivity. Meanwhile, to make sure the results are robust and reliable, we implemented Bayesian Hierarchical models on the datasets to test if the analyses on the parametric level were consistent with that from canonical ANOVA analysis. Our pre-registered ANOVA analysis with a sample size of 40 participants showed that neither Participants' beliefs nor Agent's beliefs caused any influence on participants' perceptual sensitivities (*d'*), yet there was a marginal effect from the Agent's beliefs on participants' decision criteria. But the exploratory analyses with a larger sample size (N=55) revealed a significant main effect of the Agent's belief on participants' reaction time. Meanwhile, the Bayesian analyses indicate that the probabilities of influence from Participants' beliefs on criteria were very large (P_θ =0.960), so was the effect of Agent's beliefs (P_θ =0.962). These results suggested that when no social interaction exists, participants have a weak tendency to represent other's beliefs. The results of this chapter integrated various behavioural indices and thus avoided the limitation of earlier studies which mainly drew their conclusion on a single dependent variable. It included Bayesian analysis to guarantee the conclusion is solid, showing the potential of using novel testing and analyzing approaches in studying mentalizing.

In Chapter 3, we directly tested the 'agent-object' hypothesis: does including agent-object interaction encourage participants to represent other's beliefs? To explore this question, we created two social contexts and filmed real-life videos to increase the study's ecological validity where a female actor interacted with a toy robot or not. In one condition, the actor expressed that the toy was not hers and she's reluctant to play with it. As a result, in the experimental videos, she showed no goal-directed movements such as pushing or reaching, and remained as a mere observer, as in Chapter 2. On the contrary, in the interactive condition, she stated in the introductory video that the toy belongs to her and she was going to show a demo of the game she used to play with the toy. In the experimental videos, the actor then had the pushing and reaching for the toy movement to indicate an interactive intention. In both conditions, participants were required to report whether they saw the near-threshold Gabor pattern on the robot's helmet or not at the end of each video and the signal detection approach was adopted as in Chapter 2. We then analyzed participants' perceptual sensitivities, decision criteria and reaction time under each condition. Results in the non-interactive condition replicated our findings in Chapter 2, where a weak influence from the Agent's beliefs on participants' decision criteria was found. However, the more interactive social context we created didn't make participants more involved in processing other's beliefs, and this is highly likely due to the inappropriate timing of the reaching behaviour at the end of each video.

From Chapter 4, we switched to investigate whether the 'self-agent' interaction can change participants' propensity to take another's visual perspective, especially when they are confronted with multiple conflicting perspectives. We hypothesized that the social information or perceptual features of the agent can change how we humanize them, which will impact on the 'self-agent' relationship and make a change to mentalizing. To address this question, we created a Virtual Reality scenario, with two virtual agents and the participants sitting around a table. Participants needed to identify canonical or mirror-reversed letters oriented to different directions. In Experiment 1, we manipulated the agents' social identity, whether they were ingroup or outgroup members to the participant, and in Experiment 2, we altered the agents' perceptual features such as being a moving or a statue-like agent. In both experiments, we found that participants were more prepared to take the more humanised agent's perspective, and letters oriented to these agents were reported more accurately or more quickly. Such results suggest that the humanisation process might be the gateway towards VPT.

Following up on Chapter 4, in Chapter 5 we investigated the neural mechanism underlying the bias to take a more humanised agent's perspective. We adopted the ingroup/ outgroup setting in Chapter 4, and used fNIRS to measure neural activities when participants were taking the ingroup, or the outgroup member's perspective, or when they need to select a perspective to take when both agents were present. Participants wore the VR headset together with the fNIRS equipment when doing the task, and we recorded brain activities from bilateral temproparietal junctions and its adjacent areas. The behavioural results replicated our findings in Chapter 4. For the neuroimaging results, we found that brain areas such as TPJ, pSTS, and IFG are involved in the ingroup bias effect during the perspective taking task. However, no area was specifically active when solving the perspective-selection problem.

6.2 Links between current results with previous mentalizing models

To uncover the mechanism underlying the mentalizing process, researchers have put forward various models. In particular, Apperly and Butterfill (2009) posited the two-system account to address the tension between cognitive flexibility and efficiency during mentalizing. In his theory, the implicit system takes charge of those mentalizing tasks which require simple and fast social responses based on other's mental contents, while the explicit system requires deliberate thinking, thus is usually responsible for offline reasoning of other's mental states. According to Apperly's theory (2009), as the efficiency system is rigidly constrained in order to cope with fast-changing situations, the content it can represent is rather superficial thus it can offer quite limited information for the explicit system. Consequently, Apperly argued that in most cases the two systems operate in parallel with little information exchange.

Other models focus on how people represent other's mental states in order to predict their future social behaviours. For example, in a series of studies, Tamir and Thorton pointed out that people are likely to represent other's mental states on three dimensions: rationality, social impact and valence (Tamir & Thornton, 2018; Tamir, Thornton, Contreras, & Mitchell, 2016; M. A. Thornton, Weaverdyck, Mildner, & Tamir, 2019; X. M. A. Thornton, Weaverdyck, & Tamir, 2019). Their interpretation of other's current mental states can largely be explained by these three orthogonal dimensions. Interestingly, for agents more similar to themselves, they seem to evaluate their mental states more distinctly than those of strangers. Conway et al. (2019) agree that people use dimensional representations to process other's mental states. They raised a mind-space model, and further pointed out that various social context would influence how people represent another's mental contents in the 'mind-space' and thus change their predictions on another's behaviour.

Other literature claims we should investigate real-life mentalizing issues (Bohl & van den Bos, 2012; Christensen & Michael, 2016). Based on Apperly's two-system model, Bohl and van den Bos (2012) further put forward that for most real-life mentalizing tasks, the two systems should work simultaneously and social behaviours should be achieved by the interaction between them. Critically, they stressed that the current ToM approaches have a common blind spot which is the neglect of the enabling factors for social interaction, such as the contextual and environmental factors. Christensen and Michael (2016) proposed that instead of having two systems addressing mentalizing phenomena, real-life mentalizing tasks might have one specialized system which is able to represent all belief attribution facts, yet to explicitly use such information requires cooperation from a range of other modules, such as the face-processing system or the spatial perception system.

In this current thesis, we were not aiming to evaluate the feasibility of these different models, yet we would like to attend to the blind spot in previous research, that how people start to engage in mentalizing. That is, are there any factors could change people's propensity to involve in thinking about other's mind? The answers to this question have been hinted in previous research, such as Conway et al. (2019) reported that the description of an agent to be 'oversensitive' might change participants' predictions on the agent's next behaviour, and Tamir found that people's interpretation of a close other's mental states is more diverse than that for a nonspecific agent. Such results suggest the propensity that people engage in mentalizing may rely on the social context and the 'self-other' relationship. However, no study has come up with a comprehensive model and tested those hypotheses systematically. As a move towards these goals, we constructed the mentalizing triangle model first by distinguishing the key elements in these processes. Mentalizing is a process of a participant making inferences about another's mental state, most commonly with respect to an object (at least in cognitive tasks). Thus, self, agent and object together constitute the basic elements in these processes. These three elements can interact with each other; hence they construct a triangle model (.). I propose in this thesis that these relationships are key to how people engage in mentalizing. That is, each pair of these relationships can change in various contexts, and they act as gateways to decide if a person mentalizing on others or not. When there is no relationship in the triangle, people are less likely to consider what others are thinking, as there are few reasons to do so. In contrast, when self, agent and object are closely related to each other it is highly likely that mentalizing is spontaneous. For instance, we all experienced that when choosing a gift for someone we care, we may involuntarily attend to products in the shop window and evaluate if the person would like them or not.

In this thesis, I examined the 'agent-object' and 'self-agent' hypotheses in the triangle model. The 'agent-object' relationship is usually conveyed by background information or goal-directed actions. They are the social cues which offer us evidence based on which we can infer how the agent is related to the object. We tested the 'agent-object' hypothesis with a newly designed false belief task (FBT). Previously when researchers implemented this task, the agent would have some social interactions with the object, whereas experiments with fewer such social cues (e.g. only show the interactive movements in the introductory videos, see Schneider, Slaughter, & Dux, 2017) tended to report smaller effect. Based on these previous results, we first measured the baseline effect when no social cues were present, i.e. the agent was a mere observer. Then in a second study, we compared participants' performance in conditions with and without the social interaction between the agent and the object. In contrast to our hypothesis, adding social interactions between the agent and the target object didn't encourage participants to be more involved in spontaneous mentalizing. Based on a post-hoc inspection of the data, we suggest that the odd pattern of results may have occurred because the 'reaching' action at the end of the video was distracting and might make participants falsely assume that the agents knew the robot was always behind. Thus, the agent's beliefs had little influence on the results in the interactive condition. Unfortunately, this means we cannot draw any strong conclusions about the agent-object side of the mentalising triangle.

To test the 'self-agent' hypothesis, we designed a novel perspective selection task in Virtual Reality, and manipulated the information of the agent to the participant. Our hypotheses are based on the humanization account, which has shown that people tend to humanize others differently in life, and the extent to which we humanize others can bring distinct social cognition results. Through two behavioural experiments and one neuroimaging study, we found that individuals are better at processing the visual contents of an ingroup agent, or a moving agent, as they were more accurate or faster when identifying items oriented to these agents. It is difficult for us to explain such a result with the two-system account. Based on this account, perspective-taking in our study is more likely to be implicit as no explicit instruction on perspective-taking is given, and we even didn't mention the word 'perspective' in the task. Participants only have 3 s to respond in each trial, thus it is less likely that they put a lot of time in mentalizing about the agent. According to Butterfill and Apperly (2009), the implicit mentalizing process is usually seen as an obligatory process and cannot be altered by contextual factors. Other researchers even treat is as a rapid system which is likely to be involuntary (Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010) and suggest it is impervious to a number of social or non-social factors (Cane, Ferguson, & Apperly, 2017). But in these studies, we repeatedly found that agents on different humanisation levels invoke participants to treat their perspectives differently, and there is a robust advantage for the more humanized agent's perspective. Hence, we believe before the mentalizing system is initiated, people may first process the 'self-agent' information, that agents are more similar to us are more humanised, then we tend to engage more in processing their social information including their perspectives.

Beyond these behavioural results, we also explored neuroimaging evidence for the ingroup advantage in perspective-taking and perspective-selection. In line with ample previous evidence, our results indicate that brain areas previously found to be related to mentalizing, such as the right temporal-parietal junction (rTPJ), are also involved in selectively processing other's mental contents. Other brain areas, for example the superior temporal sulcus (STS) and the inferior frontal gyrus (IFG), which activate in a series of social cognitive tasks such as ingroup favouritism and biological motion, were also involved in advantageously taking the ingroup agent's perspective. Such results further support the claim that the mentalizing process differs when people are interacting with different agents, thus supporting the 'self-agent' hypothesis in the mentalizing triangle. In addition, it also hinted a link between mentalizing and other core social cognitive processes, and future studies may further investigate how the ingroup advantage in perspective-taking can be related to other social cognitive domains. For example, a meaningful question is related to the causal relationship between ingroup advantage in perspective taking and ingroup favouritism. As perspective-taking is an initial step for people to process other's social information, it is quite likely that such bias in perspective selection will have prolonged effect on how people attribute minds, empathize others or form stereotypes, thus enrich our understanding in how intergroup relationship is formed and develops.

6.3 Methodological implications

This thesis highlights the significance of using several novel methods in investigating the complicated mentalizing problem. Firstly, the signal detection theory (SDT) has been shown to be a useful tool to provide more information when people are inferring another's mental content. Previous studies usually rely on reaction time and accuracy, and hypothesize that people should be quicker and more accurate when they need to respond to more familiar or more expected stimuli. However, the index of accuracy is often subject to ceiling or flooring effect, and reaction time is vulnerable to other factors apart from other's beliefs. Besides, both of them are indirect indices which cannot directly reflect the nature of participants' decision-making process. In contrast, the SDT teases out two components when people are making decisions: their perceptual sensitivity and decision criteria. Compared with decision criteria, the perceptual sensitivity is a more stable index and are usually not influenced by external factors such as reward or probabilities of signal or noise, but is subject to expectational or attentional factors. The decision criteria are more sensitive and more changeable with contextual factors. Therefore, with these two indices, we are able to tell if other people's mental contents change participant's perceptual ability or not.

The results from Chapter 2 and 3 also suggest that investigators can implement Bayesian Analysis in exploring social cognitive issues. Bayesian Hierarchical Analysis allows researchers to make the most use of all datasets, as it weighs less on the extreme values. Moreover, after iterating analyses at least thousands of times, it can tell researchers how independent variables interact with each other on the parametric level, thus the conclusion is more robust compared with traditional analysis such as ANOVA.

Our studies also showed that VR has great potential in addressing social interaction problems. Compared with canonical psychological methods which usually use static pictures in experiments, VR allows us to present social stimuli in a way that is more like real life. Critically, it allows us to manipulate the dynamic features of the virtual avatars, thus we are able to examine our hypotheses in a more ecological however controllable manner. With more portable devices being designed, future studies may be able to provide participants more immersive feelings, and record data from more modalities, e.g. collecting participants' movement data using motion trackers. Such improvements will largely benefit investigators to study social cognition to a new level and breed more inclusive and testable models.

The last experimental study indicates fNIRS as an ideal tool to study social interaction. As listed above, fNIRS has a high tolerance to participants’ movements, therefore it maintained the natural feeling when participants are involving in social interaction. Its relatively high temporal resolution also allows researchers to conduct event-related analysis, to reveal more robust brain signals related to the ongoing tasks. Besides, with more portable fNIRS devices, participants are able to navigate the natural world while their brain signals are recorded, therefore it will largely benefit researchers who want to relate their neuroimaging findings to situations in real life.

In general, our studies shed light on how to implement new research or analysis approaches to increase the validity of social cognitive studies. Beyond the points stated in the previous paragraphs, we consider that pre-registration should be a practice to pursue in future studies, as it can better guarantee the objectivity of the results and promote more thorough thinking before the experimental design.

6.4 Limitations of the current study and future directions

There are several limitations to the current thesis. The improper timing of the ‘reaching’ action in Chapter 3 largely limited our interpretation of the result of that experiment. Future studies may consider moving this movement into the end of each video, to rule out the alternative explanation of the attentional cue effect. To ensure the reliability of our results, it is also necessary if we balance the appearance of the toy robot between the interactive and noninteractive conditions.

Another limitation is the lack of measurement on the prefrontal cortex in Chapter 5. With more ideal VR headset, we would be able to reveal how prefrontal areas respond to perspective selection in our current task and to scenarios where participants need to interact with others.

The last limitation is the lack of testing the ‘self-object’ relationship hypothesis. Many previous studies have hinted that people treat self-relevant objects differently in cognition. In a novel study, Truong discovered that newly acquired ownership can modulate the prior-entry effect when participants were asked to remember an array of ‘self-owned’ or ‘other-owned’ items (Truong, Roberts, & Todd, 2017). It would be intriguing to test whether an object which has a high self-relevance would also attract more attention and trigger the spontaneous mentalizing process. Future studies may also consider participant’s preference of objects when designing experiments.

In relation to the general picture of future research on perspective taking and theory-of-mind, I believe that researchers might benefit from more ecological paradigms. As laboratory paradigms tended to use repetitive trials to obtain robust result patterns, participants might soon lose interest in the experimental task therefore their performance may not represent their reactions in real life. A way to overcome this shortcoming could be designing storytelling trials, where trials in a session are different from each other, but they connect and form a comprehensive story. For example, in the false-believe task, the object can change its location in different hiding spots in a room, instead of only between two locations in the classic paradigm, and by manipulation the actors ‘seeing-knowing’ behaviours we can test participants knowledge towards the actor’s belief at different time points. In this way, we can better maintain participants’ attention to the task. Besides, we can also further exploit the advantage of VR and design certain tasks where participants can interact with the virtual agent and react based on their understanding of the agent’s belief. In this way, we may observe more natural responses from participants and better predict their behaviours in life.

6.5 Conclusion of the thesis

In this thesis, we examined whether a mentalizing triangle model, composed of self, agent and object can explain how mentalizing processes such as theory-of-mind (ToM) and visual perspective taking (VPT) operates. We propose the ‘self-object’, ‘agent-object’ and ‘self-agent’ relationships act as gateways through which we engage in mentalizing. We tested the ‘agent-object’ hypothesis in ToM, and found participants have a weak trend to represent other’s beliefs in the minimal social environment. But our current manipulations on the ‘agent-object’ interaction couldn’t boost spontaneous mentalizing. We then tested the ‘self-agent’ hypothesis with a VPT task. Across three experiments, we provided evidence that participants prefer to take a more humanised agent’s perspective. Neuroimaging results further showed the bias to take an ingroup agent’s perspective may be related to the rTPJ and IFG area. This thesis includes studies that implemented a series of novel research or analysis approaches to guarantee the results are robust. The mentalizing triangle model offers a new direction to understanding mentalizing mechanisms, and shed lights on future research orientation.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Bohl, V., & van den Bos, W. (2012). Towards an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Frontiers in Human Neuroscience*, *6*(SEPTEMBER), 1–15. https://doi.org/10.3389/fnhum.2012.00274

Cane, J. E., Ferguson, H. J., & Apperly, I. A. (2017). Using perspective to resolve reference: The impact of cognitive load and motivation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(4), 591–610. https://doi.org/10.1037/xlm0000345

Christensen, W., & Michael, J. (2016). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, *40*, 48–64. https://doi.org/10.1016/j.newideapsych.2015.01.003

Conway, J. R., Coll, M. P., Cuve, H. C., Koletsi, S., Bronitt, N., Catmur, C., & Bird, G. (2019). Understanding How Minds Vary Relates to Skill in Inferring Mental States, Personality, and Intelligence. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000704

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & 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. https://doi.org/10.1037/a0018729

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition*, *162*, 27–31. https://doi.org/10.1016/j.cognition.2017.01.018

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of*

America, *113*(1), 194–199. https://doi.org/10.1073/pnas.1511905112

Thornton, M. A., Weaverdyck, M. E., Mildner, J. N., & Tamir, D. I. (2019). People represent their own mental states more distinctly than those of others. *Nature Communications*, *10*(1), 1–9. https://doi.org/10.1038/s41467-019-10083-6

Thornton, X. M. A., Weaverdyck, M. E., & Tamir, X. I. (2019). The social brain automatically predicts others' future mental states. *Journal of Neuroscience*, *39*(1), 140–148. https://doi.org/10.1523/JNEUROSCI.1431-18.2018

Truong, G., Roberts, K. H., & Todd, R. M. (2017). I saw mine first: A prior-entry effect for newly acquired ownership. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 192–205. https://doi.org/10.1037/xhp0000295