SOCIAL AND CULTURAL CONTRIBUTIONS TO METACOGNITION

Elisa van der Plas

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
of
University College London.

Wellcome Centre for Human Neuroimaging, Institute of Neurology
University College London

May 6, 2022

I, Elisa van der Plas, 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 work.

Abstract

Explicit metacognition is a hallmark of human consciousness. Its central role in the exchange of knowledge within social groups suggests that it may be shaped by social interactions. But whether, and how, social interactions may exert an effect on metacognition remains unknown. The experiments conducted in my PhD exemplify each in their own way how metacognitive ability is related to the ability to understand other people's minds (mentalizing). Chapter Two and Three show that people with compromised mentalizing ability are also more likely to have metacognitive difficulties. Contrary to the common belief that people have privileged access to their own mental states, I found that people infer their mental states indirectly from their behaviour—similar to how they infer the mental states of others. Correspondingly, people who are unable make such inferences about others (as is the case in Autism Spectrum Condition or ASC) also tend to have difficulties with doing so about themselves. Chapter Four and Five show that cultural differences in collaboration and interaction affect metacognitive ability. Across two studies, I found that Chinese students had better awareness of their own and others' mental states than occupation, age, income, gender and performance matched English students. This enhanced ability to process new evidence and correct errors generalized to how the different populations processed new social advice. Together, this work suggest that metacognition is deeply rooted in social interaction and culture.

Impact Statement

How do humans learn to be conscious? Do people have privileged and direct access to their own minds, or do we infer our own thoughts and feelings, just like we infer the mental states of others? During my PhD I studied these questions—which have attracted widespread attention across the fields of philosophy, psychology and neuroscience for centuries—with innovative behavioural testing and computational modelling of cognitive processes. Explicit metacognition, the set of cognitive processes involved in "thinking about thinking", allows for efficient self-control, such as changing your mind when you are wrong; and facilitates social communication, such as convincing others to change their mind when they are wrong. The central role of explicit cognition in social interaction suggests that it may rely on similar neurocomputational mechanisms to social cognition. The results presented in this thesis support this hypothesis, and further show that explicit metacognition is socially malleable. This thesis advances the field in two ways. From a theoretical point of view, it bridges two largely independent studies of consciousness—allowing each discipline to adopt insights from the other. For example, difficulties in understanding others' minds is often said to be a core feature of Autism Spectrum Condition (ASC), but my results suggest that metacognition plays an equally important role in this condition. From a practical perspective, understanding which contexts facilitate people to have more realistic self-views may inspire new approaches to contemporary societal issues, such as the spread of misinformation and overconfidence among leaders. A combination of metacognition and mentalizing difficulties may be at the root of some of these problems, but the social malleability of these processes suggest that a solution may be more achievable than previously thought.

Acknowledgements

Doing a PhD has been eventful and educative. I would like to thank the supervisors who have assisted me through this period—Steve Fleming, Dan Bang and Anthony David—thank you, you never failed to guide and advise me. The cross-cultural project was a special project in that it involved recruiting data in Beijing, China. I enjoyed working with Nick Wright, Shiqi Zhang and Jian Li, and particularly enjoyed guiding Keer Dong towards her MSc degree. Special thanks to David Mason and Franky Happé for having collaborated with me, it was a pleasure to learn from you.

At my very first MHJ colloquium I met my role model and friend Nuala. Working with you and David has been an intellectual paradise. Tom, I couldn't have met a better person during induction week! You always understand me and discovering London with you is so much fun. At a personal level my heart goes to Cyril and my family (including my sisters, parents and ”sister” Gabriella), you taught me a lot about love and where I should set my priorities in life. Those priorities are right there, where you are.

Contents

Chapter 1

General Introduction

1.1 Understanding minds

Explicit metacognition, the conscious evaluation of other cognitive states and processes [Dunlosky and Metcalfe, 2009, Flavell, 1979, Nelson, 1990, Shea et al., 2014] is a cornerstone of our species' success [Frith and Frith, 2003]. Explicit metacognition allows humans to express frustration and anger with words rather than actions—a skill that promoted our survival over other, more powerful, animals. The ability to understand minds is essential for many everyday situations, such as being able to recognize hesitation in someone's consent; or the ability to appreciate and evaluate one's own feelings. It is not surprising, therefore, that research has thoroughly investigated both *metacognition*, thinking about one's own mind; and *mentalizing* the sort of metacognition that involves thinking about other people's minds. The link between both processes has been debated and is a central theme of this thesis.

1.2 Explicit metacognition

A canonical aspect of metacognition that has been widely studied in laboratory tasks and educational settings is the ability to reflect on the question "Are you sure?" regarding our own knowledge, perceptions and decisions [Flavell, 1977, Nelson, 1984, Shea et al., 2014]. A person is considered to have "better" metacognition when their confidence is more tightly coupled to their actual performance, i.e., they report higher confidence when they are right, and lower confidence when they are wrong. Explicit metacognition refers to instances in which people are explicitly probed to report on an internal belief (e.g., expressing confidence), whilst implicit markers are usually inferred from their behaviour on a task (e.g., response times). Even if explicit and implicit forms of metacognition are usually correlated, there are striking reports of people behaviourally adjusting their actions to their accuracy, but not being consciously aware of doing so [Gazzaniga, 1995, Gazzaniga et al., 1962, Nelson, 1984].

Having good explicit metacognition facilitates navigating everyday life in a number of ways. First, it allows for more efficient self-control, such as asking for someone's advice when you realize that you do not know the answer. This point is particularly pertinent to educational settings, where children with poorer metacognition may not notice their need for help and miss out on instruction or support from teachers or parents [Bakracevic Vukman and Licardo, 2010, Sternberg, 1998].

Second, explicit metacognition facilitates social interaction and collaboration, such as reliably expressing how confident you are when you are advising someone else [Bahrami et al., 2010, Bang et al., 2017]. This is relevant, as collective decisions tend to be better when team members can more reliably indicate whether their own opinion should be adopted by the rest of the team or not [Bahrami et al., 2010, Bang et al., 2014, Bang et al., 2017, Fusaroli et al., 2012]. A study on leaders from U.S. based firms showed that managers, who tend to have more impact on the decisions made by other employers, are more conducive to the firm's success when they have better metacognitive efficiency [Cho and Linderman, 2019].

Third, the ability for metacognition has been associated with a range of mental health symptomologies. One study on a large general population sample [N=995] found that both excess over- and under- confidence and metacognitive efficiency are associated with different extremes of psychiatric symptoms. Whereas low confidence and better metacognition are associated with anxiety and depression symptoms; high confidence and poorer metacognition are associated with compulsive behaviours and intrusive thoughts [Rouault et al., 2018]. This suggests that maintaining the right level of self-insight is essential, not only for one's social interactions and vocational success, but also to maintain a healthy mental wellbeing.

1.2.1 Explicit mentalizing

The general set of processes involved in making inferences about other peoples' preferences and mental states is referred to as mentalizing or Theory of Mind (ToM; [Baron-Cohen et al., 1985, Baron-Cohen et al., 1986, Frith and Happé, 1999]). Mentalizing can be tested as the ability to understand that a fairy tale protagonist lacks some knowledge that the reader themself has, e.g., "Little Red Riding Hood is not afraid of the wicked wolf because she believes it is her grandmother!" [Wimmer and Perner, 1983]. In this case, a child is considered to "pass" a *false belief test* when they are able to separate their own beliefs and knowledge from those of others. Typically, children start to pass false belief around the age 4-5 (for a meta-analysis on this topic, see: [Wellman and Liu, 2004]).

Like metacognition, mentalizing has been shown to be beneficial for everyday functioning. Mentalizing protects us from engaging in antisocial behaviour [McGauley et al., 2011, Woodcock et al., 2020], vulnerability to social deception [Baron-Cohen et al., 1992, Sodian and Frith, 1992, Yirmiya et al., 1996b] and is positively correlated with relationship satisfaction rates [Cahill et al., 2020]. In addition, mentalizing disruptions have been associated with schizophrenia [Achával et al., 2010, Hooker et al., 2011, Martino et al., 2007] and autism spectrum condition or ASC [Happé, 1994, Jolliffe and Baron-Cohen, 1999, Spek and Wouters, 2010, White et al., 2009] (for a meta-analysis, see: [Chung et al., 2014]). This suggests that, just like metacognition, mentalizing is essential for navigating social interactions and everyday life, and that disrupted mentalizing efficiency is associated with a number of mental health conditions.

1.3 Theories about the relationship between metacognition and mentalizing

In the previous section I introduced the concept of explicit metacognition, a form of thinking about one's own thoughts; and explicit mentalizing, a form of thinking about another person's thoughts. Both these forms of cognition are important to sustain a healthy wellbeing and for navigating daily life. In this section, I will discuss the relationship between both forms of cognition and how each may be installed over the course of a child's development.

One view in the field is that metacognition and mentalizing share one common *meta-representational faculty* [Frith and Frith, 2012, Ryle, 1949] [Dimaggio and Lysaker, 2015, Fleming and Daw, 2017] which is thought to work separately from "first-order" forms of cognition, such as reward learning and perception. Carruthers takes a prominent philosophical position in this debate by arguing that the capacity for metacognition results from turning our understanding of other people's mental states upon ourselves [Gazzaniga, 1995, Gazzaniga, 2000, Gopnik, 1993, Wegner, 2002, Wilson, 2002]. This *"Mentalizing-is-prior"* hypothesis posits that there is one meta-representational faculty that receives input from how other people behave and feel, which is then coupled to self-referential information about how oneself behaves and feels.

One testable prediction from this hypothesis is that the mindreading shortcuts or rules one learns by observing others also serve to understand the self. In other words, Carruthers' view suggests there is no such thing as privileged access to one's own mind; there is only an indirect inferential process which involves applying mindreading rules to oneself, e.g., "Because those around me usually report high confidence when they respond quickly, I can assume from my fast responding that I am confident" [Carruthers, 2009]. A direct prediction that follows from this is that, to the extent that children are exposed to talking about feelings, they may become better at understanding other people's minds, which then in turn may be turned inwards to improve their metacognitive efficiency.

Some empirical evidence seems to be in line with the idea that people do not engage in direct introspection, but instead "read out" their intentions from past behaviour or the situation they find themselves in. Experimentally manipulating response times causally affect the confidence level people attribute to themselves and others [Patel et al., 2012], suggesting that people use response times as a proxy of confidence instead of directly introspecting a feeling of confidence itself. Furthermore, a large field of research suggests that people infer the motivations for their decisions from the situation they find themselves in at the time of a commitment (even if the situation is experimentally decoupled from the decision itself [Gazzaniga, 1983, Gazzaniga et al., 1962, Gazzaniga et al., 1977, Nelson, 1984]), suggesting that people cannot or do not introspect their own motivations directly but infer these indirectly from external cues.

A related theorized relationship between metacognition and mentalizing is the *"Metacognition-is-prior"* hypothesis, which is the direct opposite of the mentalizing-is-prior hypothesis. This theory suggests that people do use introspection, and specifically, that introspection allows people to appreciate from what they feel how others must be feeling [Goldman, 1993, Goldman and de Vignemont, 2009]. In other words, only through imagining what it would personally feel like to be in another person's shoes can people imagine what others feel. This hypothesis also makes a number of testable predictions. It predicts, just as the mentalizing-is-prior theory, that metacognitive and mentalizing efficiency are associated. But it also predicts, in contrast to the mentalizing-is-prior theory, that people do not use their own behaviour to infer their own mental states (because they have direct access to those mental states via introspection) and that having constrained access to social communication during childhood should not hinder the development of metacognition.

A third prominent view is the *"One mechanism, two modes of access"* hypothesis, which maintains that there is one meta-representational faculty that receives two distinct sources of input: one that descends from perception (containing evidence about others) and one that descends from introspection (containing evidence about oneself). This view predicts, in line with the mentalizing-is-first hypothesis, that there exist cases of compromised introspection albeit with intact mentalizing; but no cases in which mentalizing is compromised without damaged introspection [Frith and Happé, 1999]. This is because the labels used to verbally describe the mental states of oneself or others are learned via social interactions. This hypothesis, as well as the mentalizing-is-prior theory, provides an intuitive framework that can explain a number of clinical conditions, such as ASC, as a general *mind-blindness* in which mentalizing difficulties precede metacognitive difficulties later in life [Frith and Happé, 1999, Happé, 2003].

Finally, the *"Two mechanisms"* hypothesis predicts that there are two separate mentalizing and metacognition faculties, which each build upon separate conceptual models and work on distinct types of evidence (descending from perception and introspection, respectively; [Nichols and Stich, 2003]). This idea suggests that the mental shortcuts used to understand others are not the same as those used to understand the self. In other words, this theory suggests that there should be no correlation between mentalizing and metacognitive efficiency. On the other hand, a positive correlation between mentalizing and metacognitive efficiency would be consistent with three out of the four models: the one mechanism/two modes of access model, as well as the mentalizing-is-prior and metacognition-is-prior hypotheses.

1.4 Experimental support and confounds

Given that metacognition and mentalizing have rarely been empirically investigated in tandem, it remains unclear which of these theories is best supported by the data. There are reported cases of positive correlations between mentalizing and metacognitive efficiency in the same individuals [Nicholson et al., 2020]—in favour of the first three models; but also reports where such correlation is lacking [Carpenter et al., 2019a]—in favour of the two mechanisms model. And although cases of mentalizing and metacognitive dissociations are reported in ASC [Carpenter et al., 2019b, Wojcik et al., 2013], there are also studies that report a commensurate impairment of both metacognition and mentalizing efficiency in ASC [Grainger et al., 2016a, Nicholson et al., 2020, Williams et al., 2018], in favour of the mentalizing-is-prior and the one mechanism, two modes of access theory.

One caveat of empirical research on metacognition is that metrics of mentalizing and metacognitive efficiency are often confounded by *lower-order* or *type 1* cognitive processes. These are the types of cognition that are the subject of the mentalizing or metacognitive reflection, e.g., perception, verbal fluency, learned associations, other people's behaviour. Lower-order processes can confound the measurement of metacognition and mentalizing in a number of ways, which I will explain below.

First, in the domain of metacognition, people tend to be more aware of their own performance when they are performing well (versus poorly) on a task [Maniscalco and Lau, 2012, Pouget et al., 2016]. In turn, the same person's performance awareness will drop when the task is made more difficult [Fleming and Lau, 2014, Masson and Rotello, 2009, Rahnev and Fleming, 2019] —potentially affecting metrics of metacognitive efficiency on tasks where accuracy is left to vary within and between individuals. For example, ratings indicating how confident someone is that they saw some target (a metacognitive process) may be confounded by lower-order perceptual ability, which needs to be avoided because a reliable metacognitive metric should only encompass the thought process about perception, and not perception itself.

In addition to choice accuracy, a general tendency for over- or under-confidence, called *metacognitive bias*, is a common confound in the measurement of metacognitive efficiency [Masson and Rotello, 2009, Fleming and Lau, 2014, Rahnev and Fleming, 2019], e.g., as is the case in the commonly used metric of metacognition called the Goodman-Kruskal gamma coefficient (G [Nelson, 1984]). In other words, someone can have a general tendency to rate high confidence, but while doing so, may still be able to sensibly differentiate between their errors and correct responses. A reliable index of metacognitive efficiency accommodates this dissociation.

Third, some tasks of mentalizing and metacognition provide trial-by-trial feedback (e.g., on trials where you are correct and confident you get 10p; incorrect and unconfident -10p; correct or incorrect but unconfident 0p [Nicholson et al., 2020]), which introduces the possibility that metacognitive metrics derived from such tasks are confounded by individual differences in the ability to learn from feedback. This point is particularly pertinent to studies examining relationships between metacognition and mental health symptomology, as some mental health problems are characterized by difficulties with learning from ambiguous and incomplete feedback, e.g., ASC [Broadbent and Stokes, 2013, Greene et al., 2019, Reed, 2019, Robic et al., 2015, Sapey-Triomphe et al., 2018, Zwart et al., 2018], Post-Traumatic Stress Disorder [Myers et al., 2013], Obsessive Compulsive Disorder [Becker et al., 2014] and Major Depressive Disorder [Herzallah et al., 2013].

Finally, as mentalizing tasks often require a participant to verbally describe another person's mental state [Abell et al., 2000, David et al., 2008, Livingston et al., 2019b, Rosenblau et al., 2015, White et al., 2011, White et al., 2009], metrics of mentalizing efficiency may be confounded by individual differences in verbal fluency [White et al., 2011]. It could be that people who are native English speakers are more articulate at providing answers on mentalizing tasks, leading them to score higher irrespective of their true underlying mentalizing efficiency.

1.5 Signal Detection Theory

Some of these issues can easily be resolved by making small changes to the used experimental design. For example, not providing trial-by-trial feedback on confidence ratings or using multiple choice questions instead of open responses [White et al., 2011]. But recently new analysis approaches have been proposed to resolve the caveats around the measurement of metacognitive efficiency. First and foremost, by adopting a signal detection theoretical approach to calculate metacognitive efficiency from confidence levels and decisions [Maniscalco and Lau, 2014, Maniscalco and Lau, 2012]. Signal detection theory is a model of how observers detect signals from noise. Based on the number of 'hits' (the observer correctly reports to have seen the stimulus) and the number of 'false alarms' (the observer reports to have seen a stimulus, whereas they merely saw noise) the perceptual sensitivity (d') of the observer (i.e., how well the observer can discriminate between signal and noise) is calculated in a way that is uncontaminated by differences in perceptual bias (i.e., a general tendency for an observer to report having seen the stimulus irrespective of what they saw [Green and Swets, 1966]).

It is now possible to use this framework to isolate a signal detection theory-based metric of metacognitive efficiency, called *metacognitive sensitivity* (meta-d'), which is uncontaminated by perceptual sensitivity (d') and overall confidence [Fleming and Lau, 2014, Maniscalco and Lau, 2012, Maniscalco and Lau, 2014]. Specifically, these models apply a "second-order" or type 2 signal detection framework, where detection does not consist of reporting to have seen a stimulus but reporting to be aware of errors with confidence ratings. Metacognitive sensitivity is computed from the true positive rate: of all correct trials, how often does the observer accurately report high confidence; and the false positive rate: of all incorrect trials, how often does the observer incorrectly report high confidence—making this metric independent of *metacognitive bias*, how confident or unconfident a person generally is irrespective of how much evidence they have seen [Fleming and Lau, 2014]. Importantly, when measured like this, metacognitive sensitivity (meta-d') is in the same units as perceptual sensitivity (d'), allowing for the characterization of *metacognitive efficiency* (or $Mratio$), the ratio between an observer's sensitivity to differences in their accuracy and their sensitivity to differences in perceptual stimulation:

$$Mratio = meta - d'/d' \tag{1.1}$$

Computed like this, $Mratio$ is a computational extraction of only the metacognitive aspect of perception and is, therefore, uncontaminated by differences in accuracy or average confidence. Crucially however, Mratio has been used to explore a relationship between metacognition and mentalizing efficiency only once

[Nicholson et al., 2020].

One caveat of computing metacognitive metrics using a SDT approach is that it requires a large number of trials. Recently an alternative approach to the estimation of Type 2 SDT parameters has been developed. The HMeta-d' toolbox employs Bayesian inference to sample Type 2 SDT parameters hierarchically [Fleming, 2017]. The benefit of this approach is that Bayesian inference takes into account the uncertainty of parameter estimates, such that under low trial numbers the parameter computation will be adjusted accordingly. In addition, hierarchical modelling allows for an efficient integration over both within and between-subject variability in uncertainty. For example, if two datapoints or subjects are closer together (e.g., because they are retained from the same subject or they pertain to the same clinical group) hierarchical Bayes capitalizes on statistical strength by using both pieces of information to reduce noise of the inference [Harrison et al., 2020, Kruschke, 2010].

Taken together, the few studies that have tested metacognition and mentalizing in tandem may have been confounded by methodological artefacts—making it difficult to assess theories about the relationship between metacognition and mentalizing. In chapter two, I circumvent these methodological issues by experimentally and computationally disentangling type 2 metacognitive and mentalizing efficiency from type 1 cognitive artefacts. Specifically, I compute a type 2 signal detection theoretical estimate of metacognitive efficiency and regress this variable against variability in mentalizing efficiency within the same hierarchical regression model.

In a large dataset of the general population [N=501) I find strong evidence in favour of the *mentalizing-is-prior* theory. In particular, I find that among these 501 participants those that were more capable of assessing the mental states of others were also better at recognizing their own errors. Intriguingly, I found that mentalizing was associated with a tighter coupling between response times and confidence on the metacognition task, suggesting that people do not infer their confidence directly but infer their confidence from their bodily cues, as if they were inferring the confidence of someone else. Finally, people with better social communication skills had better metacognitive efficiency, suggesting that everyday exposure to talking about feelings selectively improves the ability to understand and reflect on one's own mental states.

1.6 Meta-representational development

In the previous section, I described theoretical and empirical approaches to studying the relationship between mentalizing and metacognition. I outlined four distinct hypotheses about this relationship. Striking distinctions can be made between the mentalizing-is-prior hypothesis, which posits that metacognition develops by turning other people's descriptions of their mental states to oneself, and the metacognition-is-prior hypothesis, which posits that people rely on introspection to understand others; and the two mechanisms, one mode of access hypothesis, which posits that metacognition and mentalizing are two entirely separate processes.

One way to better understand the relationship between metacognition and mentalizing is to study their developmental trajectories. If metacognition develops by turning mindreading to oneself, one would expect the development of mentalizing to precede (or coincide) with the development of metacognition; and that this development is mediated by how much opportunity a child has had to learn mindreading rules. Instead, if metacognition and mentalizing are two separate systems, one would predict their developmental trajectory and onset to be dissimilar too. In this section I will review literature from developmental psychology on the relationship between metacognition and mentalizing.

1.6.1 Mentalizing

Mentalizing processes first start to emerge by the second year of life, along with children's first use of words that describe their mental states and desires ("want" and "know") [Bartsch and Wellman, 1995, Flavell, 1977, **?**]. The type of mentalizing that develops around this stage is called *first order mentalizing*, an understanding of (another person's) mental states, e.g., "Sally thinks that her chocolate is in the cupboard". Somewhat later, between two and a half and four years old, children also start to engage in *second order mentalizing*, an understanding of another person's mental states about another person's mental states [Leslie and Frith, 1988], e.g., "Sally thinks that her chocolate is in the cupboard whereas it is actually in the candy jar, because I saw her dad put it there when she was not looking".

There is substantial individual variability in the development of mentalizing, which has allowed researchers to assess the role of different contextual factors. One recurring environmental factor that shows to be essential for the development of mentalizing is exposure to social communication. The amount of social communication a typically developing child is exposed to—be it in the form of having more siblings [Jenkins and Astington, 1996, Perner et al., 1994], the amount of fights one has with their siblings [Dunn et al., 1991, Dunn and Brown, 1993, Jenkins and Astington, 1996] or simply the number of times people around them refer to mental state terms [Brown et al., 1996]—determine how well and early a child will pass mentalizing tests: the more they are exposed to talking about mental states, the better and earlier they pass mentalizing tests.

Insights into the development of mentalizing efficiency are also obtained by observations of children that undergo untypical mental development. Whereas 86% of typically developing four-year-old's pass second order mentalizing tests, only 20% of children with ASC pass it [Leslie, 1987, Leslie and Frith, 1988]. Strikingly, whereas children with Down Syndrome score similar to typically developing children on mentalizing tasks, deaf children born to hearing parents (but not deaf children born to deaf parents) experience mentalizing delays that are characteristic to those found in ASC [Peterson and Siegal, 1995]. Later neuroimaging research has built upon this work, by showing that deaf children's constrained access to communication has a selective impact on the development of mentalizing-specific, but not language-specific, neural regions. The *ToM network* is involved in understanding mental states and consists of the temporoparietal junction (TPJ), precuneus and medial prefrontal cortex (mPFC; [Adolphs, 2009, Kana et al., 2015]). Deaf children who have delayed access to sign language have lower activity in the ToM network when they listen to descriptions of mental states compared to deaf children who are native signers [Richardson et al., 2020], suggesting that early access to (sign) language has a direct impact on the behavioural and neural shaping of mentalizing.

In sum, this work suggests that being exposed to social descriptions of what other people feel is essential for the development of mentalizing. These results are in line with the mentalizing-is-prior and the one mechanism, two modes of access hypothesis described in the previous chapter, but not with the metacognition-is-prior hypothesis. Specifically, because the latter proposes an introspection-based precursor for metacognition that should not rely on access to other people's mental state descriptions. I will next review how this information fits in with what we know about the developmental trajectory of metacognition.

1.6.2 Metacognition

The association between confidence and accuracy that prevails the experimental characterization of metacognitive efficiency can be tested from *implicit* behavioural cues—such as 'opting-out' from choosing when one is uncertain; and from *explicit* verbal expressions, such as reporting lower confidence when one is uncertain [Frith, 2012]. Explicit metacognition becomes fully formed between the ages of 3 and 4 [Hembacher and Ghetti, 2014], although implicit precursors can be identified already among preverbal infants [Goupil and Kouider, 2019]. Intriguingly, explicit metacognition emerges around the same time as second order, but not first order, mentalizing [Lockl and Schneider, 2007, Carruthers, 2009, Lockl and Schneider, 2007], suggesting that the development of mentalizing may at least partially coincide with that of metacognition.

Individual differences in metacognitive sensitivity have been found to correlate with the structure and function of the anterior prefrontal cortex (aPFC), a brain region involved in self-evaluation [Christoff and Gabrieli, 2000, Hilgenstock et al., 2014, Shimamura and Squire, 1991, Fleming and Dolan, 2012]. This region, as well as the posterior medial frontal cortex (pMFC), precuneus and the medial prefrontal cortex (mPFC) are active during self-evaluation of performance [Fleming and Dolan, 2012], the latter regions being shared with the ToM network [Vaccaro and Fleming, 2018]. Disruptions in neural regions that are involved in both mentalizing and metacognition are predictive of neurodevelopmental conditions, such as schizophrenia [Holt et al., 2011, Modinos et al., 2011]. This suggests that the neuro-developmental basis of metacognition and mentalizing is at least partially overlapping and that these are commensurately disrupted in some neurodevelopmental conditions.

Even if functional magnetic resonance imaging (fMRI) studies of metacognitive areas in infants are uncommon, some evidence from electroencephalogram (EEG) recordings suggests that the neural development of metacognition is commensurate with that of mentalizing. Error-related negativity (ERN) is a drop of neuronal activity following errors that signals the metacognitive efficiency for *error monitoring* (realizing one's error in the absence of feedback [Charles et al., 2014]). The earliest occurrence of ERN in humans is around one year [Dehaene-Lambertz and Spelke, 2015, Goupil and Kouider, 2019], around the age that first-order mentalizing also starts to develop. Interestingly, whereas sensory brain structures become fully formed earlier in infancy, there is only a sudden increase in prefrontal myelination [Dehaene-Lambertz and Spelke, 2015, Goupil and Kouider, 2019] by the end of year one, suggesting that these regions are involved in the development of metacognitive processes too [Goupil and Kouider, 2016].

In contrast to the plethora of research on the development of mentalizing, much less is known about what contextual factors are involved in the development of metacognition [Lockl and Schneider, 2007]. Some studies suggest that the development of metacognition, just like the development of mentalizing, is conditional on access to language and social interaction [Flavell, 2000]. In other words, access to other people's descriptions of their mental states facilitates a developing understanding of one's own mental states. This metacognitive dependency on social communication may explain why metacognitive efficiency varies as a function of socio-cultural norms of communication and interaction [Proust and Fortier, 2018, Heyes et al., 2020] as well as why children's quality of relationships with their caretakers predicts their metacognitive efficiency [Roebers, 2017] (but see also: [Kim et al., 2020]). In sum, some theoretical, but very limited empirical work, suggests that access to social communication and mental states is as relevant for the development of metacognition as it is for the development of mentalizing.

The presented literature on the developmental trajectory of mentalizing and metacognition suggest that the first meta-representational ability children learn is the ability to understand what other people are thinking (first order mentalizing) which is facilitated by exposure to information about other people's mental states. At approximately the same time, children undergo a speedy development of prefrontal neural areas involved in metacognitive processing, which allow them to recognize those same mental states in themselves and engage in more complex forms of mentalizing. If this developmental trajectory of metacognition and mentalizing is true, I would expect the conditions that compromise the development of mentalizing to be commensurately obstructive to the development of metacognition.

Autistic symptomology restricts autistic infants' interest in social information as early as of 2 years old [Chawarska et al., 2003, Chawarska and Shic, 2009, Chawarska and Volkmar, 2007] (but see also: [Elsabbagh et al., 2013]). In adulthood, ASC is characterized with atypical social interest and communication, and poor mentalizing skills—perhaps limiting the amount of social information autistic children are exposed to or seek [Chevallier et al., 2012]. If mentalizing and metacognition indeed have a shared neuro-developmental trajectory, this restricted social information should have consequences for the development of metacognition too. However, it is currently unclear from the literature whether metacognitive efficiency is correlated with autistic traits of social communication; and whether autistic people have a commensurate impairment of both mentalizing and metacognitive efficiency.

In chapter two, I leverage the autistic phenotype as a testbed to investigate if access to social communication is a requisite for metacognitive efficiency. If the development of metacognition and mentalizing efficiency are commensurate, one would predict both metacognitive faculties to be compromised in autism; and that this compromised metacognitive efficiency in autistic individuals is driven by problems with social understanding and communication. To this end, I analysed the data of N=40 autistic participants and N=40 comparisons that were matched in terms of age, gender, IQ and education, and found that metacognitive efficiency was indeed compromised in autistic participants. Furthermore, I replicated the finding that metacognitive efficiency and mentalizing efficiency were correlated in the group as a whole; and that one, shared meta-representational faculty encompassing both metacognition and mentalizing could explain autistic symptomology. Together, this work suggests that autism can be considered a metarepresentational condition where restricted access to mental states may give rise to difficulties with mindreading generally—be it in reading the mind of others, or that of oneself.

1.7 Sociocultural malleability of metacognition

In the previous section I described research suggesting that the development of metacognition, including the underlying neural machinery, is affected by social communication. First order mentalizing develops by year one, and is later followed by a fast spike in myelination in frontal areas involved in second order mentalizing and metacognition, around the same time that the first behavioural indicators of complex metacognition begin to show. These findings provide evidence in favour of a mentalizing-is-prior theory, given that access to social descriptions of mental states seem to play an essential role in the development of various forms of metacognition (which would not be predicted under the metacognition-is-prior hypothesis, where metacognition is informed by introspection). If this theory is correct, and the development of metacognition is indeed dependent on access to social descriptions of mental states, one would expect differences in norms of social communication across cultural groups to have downstream effects on the development of metacognition too. In this section, I will review cross-cultural studies to understand if they can teach us something about the origins of metacognition.

A commonly used cultural distinction is that between collectivist and individualistic cultures. Cultures in South and East Asia, South America and a number of islands in Oceania are often described as more collectivist in that they traditionally emphasize harmony with others, whereas cultures in Western countries like those in Northern Europe or the United States are commonly viewed as more individualistic. In the late 19th century it was described that these cultural differences manifest in distinct political and educational styles [Cripps, 1998, Weber, 1905], and even in different styles of thought [Hofstede, 2011, Markus and Kitayama, 2010, Oyserman, 1993]. For example, more collectivist forms of education tend to focus on learning from one another and working in groups, whereas individualistic cultures tend to focus on personal responsibility for one's own educational success and working alone.

1.7.1 Mentalizing

There are a number of studies that have compared mentalizing efficiency between different cultures, which together suggests that mentalizing efficiency is susceptible to cultural differences. Samoa is a Polynesian island country in Oceania in which strict social codes and etiquette govern social life. Samoan culture deems it inappropriate to address pre-verbal infants and considers it impolite to talk about mental states. A number of studies now suggest that Samoan children pass mindreading tests much later than children from various Western cultures [Mayer and Träuble, 2015, Mayer and Träuble, 2013]. Other work suggests that, beyond culture, also within-country variability can affect mentalizing skill [Liu et al., 2008], which further complicates an inference of what cultural aspect is driving the differences in mentalizing efficiency. It could be, for example, that in societies where collectivism and social collaboration prevail, people tend to more openly express and describe what they feel to improve group performance [Bahrami et al., 2010, Bang et al., 2014, Fusaroli et al., 2012, Bang et al., 2017]—which, in turn, may benefit mentalizing efficiency.

Later work showed that the way in which mentalizing is acquired varies between Chinese and American children [Wellman and Liu, 2004] in ways that are contingent on cultural differences in how information is socially broadcasted within one's social circles [Kim et al., 2018]. This suggests that, beyond mere cultural variety in mentalizing efficiency, also the mentalizing process itself may vary along the individualist-collectivist spectrum.

These results are consistent with the hypothesis that exposure to social communication about mental states affect the onset at which mentalizing efficiency develops. One caveat is that cultural norms are not necessarily homogeneous within countries, and it remains unclear whether more collectivist societies promote or compromise the development of mentalizing.

1.7.2 Metacognition

As mentioned in chapter two, computational neuroscience has found ways to test how metacognitive evaluations come about. I will build upon this work to explore whether there are cultural differences in metacognition.

Computationally, metacognitive evaluations can be described as an extension of the so-called *drift diffusion model*, which assume that an agent will reduce uncertainty about which decision to make by sampling information about the potential benefits and disadvantages of the various available options [Gold and Shadlen, 2007]. Metacognitive evaluations are often studied in the laboratory using perceptual decision making tasks such as the random-dot motion task. Random-dot motion stimuli consist of a rapidly moving cloud of dots presented briefly on a computer screen (typically for less than a second), with a proportion of the dots moving coherently in a particular direction, whereas the remainder move randomly (e.g., left or right; [Britten et al., 1992, Kim and Shadlen, 1999]). On each trial of the task subjects are asked to decide whether the dot cloud is mostly moving in one or other direction. This process can be described by computational models that assume the brain receives noisy samples of evidence about the world (e.g., whether the dots are moving left or right), and compares these samples to an internal decision threshold [Luce, 1986, Ratcliff, 1978, Vickers, 1979]. These models can predict how long the observer takes to make a decision [Smith and Ratcliff, 2004, Wald, 1947], and even how confident they are in their choice [Kiani and Shadlen, 2009].

In particular, studies that have used DDMs to explain confidence have shown that confidence is informed by evidence that is accumulated after an initial decision has been made; the "extra" evidence that is sampled after the decision is called *post-decision evidence* and can support error monitoring [Murphy et al., 2015, Rabbitt, 1966] and changes of mind [Resulaj et al., 2009, van den Berg et al., 2016]. The general finding is that people tend to update their final judgment after seeing the new evidence (i.e., becoming more confident that they were right when new evidence confirms their choice; and less confident that they were right when new evidence disconfirms their choice), and that this updating is stronger when the new evidence is more reliable or stronger [Bronfman et al., 2015, Fleming et al., 2018] and especially so when the new evidence confirm, rather than disconfirms, the initial choice (known as 'confirmation bias'; [Talluri et al., 2018]).

The posterior medial frontal cortex (pMFC) activates when people detect that they have made a mistake and adjust their behaviour accordingly [Dehaene et al., 1994, Ridderinkhof et al., 2004]. In addition, this same region tracks the strength of post-decision evidence and signals the need for behavioural adaptation. In contrast, activity in the anterior prefrontal cortex (aPFC), a region that is also involved in metacognition, mediates the impact of new evidence on people's subjective confidence [Fleming and Dolan, 2012]. In other words, susceptibility to new evidence following errors is a computational building block of metacognitive efficiency.

The link between metacognition and new evidence processing is supported by two recent studies that measured both metacognitive sensitivity and sensitivity to post-decision evidence, and asked how these aspects were related to a personality feature known as dogmatism (measured as the extent to which individuals were accepting of conflicting views on political issues [Rollwage et al., 2020, Schulz et al., 2020]). Metacognitive sensitivity predicted the extent to which subjects integrated new evidence on the perceptual task, supporting adaptive changes of mind. In turn, both of these features of decision-making were attenuated among those with higher levels of dogmatism. This finding shows that metacognitive sensitivity may promote adaptive changes of mind when new evidence becomes available. On top of that, this study suggests similar processes may govern the processing of new evidence in both low-level perceptual discrimination tasks and the broader, more subjective decisions about topics such as politics. However, whether culture is similarly associated with metacognitive efficiency is currently unclear. Some work suggests that cultural differences selectively impact the way in which metacognitive evaluations are made (e.g., in how much people are inclined to ask for help), rather than the outcome of metacognitive process itself (e.g., in average accuracy of the eventual decision; [Kim et al., 2018, Wellman and Liu, 2004]). For example, there are no or inconsistent reports of differences in overall confidence or in how well confidence ratings dissociate correct from erroneous decisions between Western and East-Asian populations [Yates et al., 1998, Yates et al., 1989]. However, several studies have found evidence for cultural differences in susceptibility to social advice [Korn et al., 2014, Mesoudi et al., 2015, Kim et al., 2018], which is considered a social form of post-decision evidence processing [van der Plas et al., 2019].

In addition, while there appear to be no cultural differences in lower-level cognitive processes, such as associative learning between Chinese and UK participants [Wright et al., 2018] and sensory processing between Chinese and German participants [Nan et al., 2006]; studies focusing on metacognitive processes, such as integration of social feedback [Korn et al., 2014, Mesoudi et al., 2015], self-reports of anticipated surprise [Ji et al., 2001, Valenzuela et al., 2010] or metacognitive judgments of confidence [Moore et al., 2018, Stankov and Lee, 2014, Yates et al., 1998], have more consistently reported differences between people with Chinese and Western backgrounds. This suggests that metacognitive processes in the form of post-decisional evidence processing is particularly susceptible to external (social) input [Shea et al., 2014, Frith and Frith, 2012]. However, a major shortcoming of these studies is that they have not dissociated metacognition from first-order performance.

To address this issue in chapter four, I leverage performance-controlled psychophysical paradigms to compare the profiles of metacognitive judgments about task performance in demographically homogeneous datasets collected in China and the UK. These samples were matched for occupation (volunteers were students at PKU and UCL respectively), age, gender, income and IQ. The results provide evidence for selectively heightened metacognition in the Chinese participants, driven by an increase in post-decision evidence processing following error trials. Specifically, Chinese participants were more likely to acknowledge having made an error when they were presented with information that conflicted with their initial view than UK participants. This study provides intriguing evidence that metacognitive processes are formed during interactions with other members of our cultural and social groups.

1.8 Advice-taking as example of metacognition

In the previous section I have discussed preliminary evidence that metacognition is shaped via social interactions and is, therefore, malleable to socio-cultural variation.

In doing so, I built upon a recently proposed theory of how culture shapes metacognitive processes [Cleeremans et al., 2020, Heyes et al., 2020]. Empirical work described in chapter four is in line with this view and suggests that cultural differences may specifically impact how people re-evaluate past choices on the basis of new perceptual evidence. In everyday situations, however, new evidence presents itself only in limited cases in a perceptual format. Instead, when evaluating our own behaviour, it is natural to turn to the help and advice of other social agents as a source of post-decisional evidence—a process called *advice-taking*.

Beyond a number of social motivations to take advice, for example, scenarios in which people comply to fit into a social group [Kelliher et al., 2011] or under the influence of an authority figure [Milgram, 1963], I will argue that genuine advice-taking (i.e., advice-taking that engenders a privately held shift in beliefs) encompasses a form of metacognition, which allows an advisee to infer whether more evidence is needed; and a form of mentalizing, which allows an advisee to infer whether the advice at hand is reliable [De Martino et al., 2013, van der Plas et al., 2019, Folke et al., 2016]. In result, an instance of advice-taking provides both information about a person's metacognitive efficiency, i.e., the extent to which a person takes advice when they themselves are wrong versus when they are right; as well as about that same person's mentalizing efficiency, i.e., the extent to which a person takes advice when their adviser is correct versus when they are wrong.

There are three lines of evidence in support of the idea that some forms of advice-taking involve a meta-representational system. First, advice has a larger impact when the advisee is uncertain about their initial judgment [Cialdini and Goldstein, 2004, De Martino et al., 2017] and when the adviser is right [Campbell-Meiklejohn et al., 2017, Campbell-Meiklejohn et al., 2010]. These results indicate that, just as some have reframed post-decision evidence processing as a way to resolve or minimize choice uncertainty, advice can be reframed as a piece of epistemic information that allows for changes of mind when it is most beneficial. Second, while involving distinct brain regions that might depend on the input format, advice-taking, metacognition and mentalizing also involved a shared set of brain regions. Areas such as pMFC and aPFC are involved in both instances of advice-taking and metacognition [Fleming and Dolan, 2012, Fleming et al., 2018, Vaccaro and Fleming, 2018]. Third, autistic individuals, who tend to have metacognitive difficulties report experiencing difficulties with understanding if someone is deceiving them or not [Baron-Cohen et al., 1992, Sodian and Frith, 1992, Yirmiya et al., 1996a] and are more likely to take advice at face validity, without fully considering its relevance [Large et al., 2019].

The idea that social advice-taking is based on the same mechanisms as perceptual evidence processing suggests that post-decision evidence processing is impartial to the type of evidence at hand (e.g., social and non-social alike). This *domain generality* of metacognition is a recent topic of discussion and some evidence indeed suggests that metacognition works similarly across different domains. For example, studies have shown that the way in which new evidence is integrated is similar across distinct tasks (e.g., perceptual and numerical tasks [Bronfman et al., 2015, Talluri et al., 2018]). Ongoing research is now building upon this finding by examining how new evidence is integrated from both non-social and social information sources [Olsen et al., 2019, Pescetelli and Yeung, 2021].

Even if previous work has provided preliminary evidence in favour of a similar processing of social and non-social sources of evidence [Behrens et al., 2008], and other work has found evidence in favour of enhanced social compliance in Chinese versus Western populations [Korn et al., 2014, Mesoudi et al., 2015], the issue of cultural differences in metacognition is still unresolved because of potential differences in the reliability of advice. In particular, folk psychology tells us that social advice is in some ways more ambiguous that other types of evidence. Whereas the reliability of one's own opinion can be readily accessed via introspection (or at least it feels like it can), the reliability of another person's view needs to be inferred by integrating over both what the adviser says that their conviction is (their confidence rating) as well as their metacognitive reliability (how well their confidence tracks their objective accuracy—a form of mentalizing [van der Plas et al., 2019]). Previous work has outlined that tracking and learning the reliability of advice is an important aspect of advice-taking [Sniezek and Van Swol, 2001, Behrens et al., 2008, Campbell-Meiklejohn et al., 2010, Gomez-Beldarrain et al., 2004]. However, how metacognition and mentalizing interact in the process of advice-taking is still unknown.

If metacognitive efficiency results from turning insights about other people to oneself, as would be predicted by the mentalizing-is-prior theory, inference about the reliability of an adviser should work similarly to making inferences about the reliability of oneself. In addition, if mentalizing and metacognition share a common meta-representational mechanism, as the results from chapter two suggest, one would expect a cultural benefit for metacognition to generalize to cases of mentalizing. In chapter five, I test this final assumption of the mentalizing-is-prior hypothesis by recruiting two matched samples of Chinese and UK populations. I extract two main variables of interest from advice-taking behaviour. First, in line with chapter four, I quantify metacognitive efficiency as the extent to which an individual is susceptible to new advice when one is wrong (and not correct). Second, even if this is not necessarily the standard use of the term, I will interpret susceptibility to advice when the adviser is correct (and not wrong) an instance of mentalizing efficiency. The reason for calling it mentalizing efficiency, rather than mentalizing ability, is that this metric is not confounded by the participant's nor the adviser's first-order performance (similar to the way in which metacognitive efficiency is not confounded by first-order performance). The way in which I achieve this separation between second- and first-order cognition in the computation of mentalizing efficiency is by generating the advice with a computational model that mimics the perceptual process of the participant. This way, I created an instance of mentalizing efficiency that is not confounded by first-order differences in the advisers' or one's own accuracy and confidence.

Consistent with the hypothesis that metacognition and mentalizing share a common meta-representational mechanism, I found that the Chinese participants' benefit in processing new evidence in adjusting ongoing errors generalized to evidence of the social type. In addition to replicating the metacognitive improvement in post-decision evidence processing of Chinese versus UK populations (chapter four), I found that the Chinese participants' susceptibility to social advice was restricted to cases in which the advice was reliable, suggesting a similar cultural susceptibility of metacognition and mentalizing.

1.9 Open questions

1. Developmental research suggests that understanding one's own mind and understanding other people's minds may share a computational basis. One difficulty with testing this hypothesis is that metrics of metacognition and mentalizing are often confounded by the first-order cognition that it evaluates. In *chapter two* I develop an experimental framework that overcomes some of these issues. While all computational models are likely to be wrong in some ways, and necessary oversimplifications, testing which of Carruthers' models most plausibly explains our observed variability between metacognition and mentalizing efficiency can provide a useful starting point from which more nuanced insights can be obtained.

2. According to Carruthers' models, the ability to understand one's own mind develops from an evolving understanding of other people's minds. If the mentalizing-is-prior theory is true, one would predict those with limited access to mental information during their development (e.g. ASC) to have greater difficulty with metacognition as well as mentalizing in adulthood. On the other hand, if the development of metacognition is independent of the development of mentalizing, as follows from both the metacognition-is-prior as well as by the two mechanisms / one mode of access hypothesis, I would predict compromised mentalizing (but not metacognitive) efficiency in autis-

tic compared to non-autistic people. To test these hypotheses in *chapter three* I measure metacognitive and mentalizing efficiency in forty autistic persons and forty gender, age, IQ and education matched comparison participants.

3. Do socio-cultural differences affect metacognitive efficiency? In *chapter four* I will present results from a cross-cultural collaboration with Peking University in Beijing that helps differentiate between the metacognition- and mentalizing-is-prior theory. In particular, if norms of collaboration promote an openness to talk about mental states, this could also engender differences in the adaptive use of new evidence to recognize and correct one's errors, as follows from the mentalizing-is-prior hypothesis. Instead, if the metacognition-is-prior theory is correct, cultural norms of collaboration should not affect metacognitive efficiency, as this theory predicts that metacognitive efficiency is informed by direct introspection rather than mentalizing.

4. So far, I have tested mentalizing and metacognition via two separate test batteries. But how do those two processes interact when they need to be applied in the same decision context, e.g., integrating the reliability of one's own perceptual decision with the advice of another person about the same perceptual problem. This question captures a final implication of the mentalizing-is-prior theory, namely, that both metacognition and mentalizing share similar computational processes. In *chapter five* I explore how and whether metacognition facilitates an ability for successful advice-taking, and whether advice-taking skill is better among those that were raised in more collectivist societies.

Chapter 2

Computations of confidence are modulated by mentalizing efficiency

"The sorts of things that I can find out about myself are the same as the sorts of things that I can find out about other people, and the methods of finding them out are much the same."
– G. Ryle in The Concept of Mind (1949)

2.1 Introduction

In 1949, Ryle famously proposed that the cognitive mechanisms employed to understand ourselves are similar to those involved in understanding the feelings and experiences of other people [Ryle, 1949]. Since then, various proposals have echoed Ryle in suggesting that *explicit metacognition*—the capacity for conscious evaluation of one's own mental states [Fleming and Lau, 2014, Frith and Frith, 2012, Fleming et al., 2010, Yeung and Summerfield, 2012] and *mentalizing*—the capacity to evaluate and understand other people's mental states [Abell et al., 2000, David et al., 2008, Livingston et al., 2019b, Rosenblau et al., 2015, White et al., 2009, White et al., 2011] have a common computational basis [Carruthers, 2009, Dimaggio and Lysaker, 2015, Fleming and Daw, 2017, Frith and Frith, 2012].

According to recent perspectives on the developmental trajectory of metacognition, while ”core” or implicit mechanisms for self-monitoring and tracking uncertainty may be in place early in infancy [Goupil and Kouider, 2019], explicit metacognition emerges around the ages of 2-3 (e.g. [Hembacher and Ghetti, 2014], see [Goupil and Kouider, 2016] for a review), and continues to be shaped in childhood and adolescence [Fandakova et al., 2017, Weil et al., 2013]. One potential driver of this continued development of explicit metacognition is that a growing understanding of other people's mental states may be used to refine awareness of ourselves [Carruthers, 2009]. For example, repeatedly perceiving a parent expressing uncertainty together with their hesitation may allow a child to recognize and express uncertainty when they themselves are hesitating. This hypothesis predicts that introspection is not a distinct natural kind, but is instead grounded in the same processes used to understand the mental states of others [Carruthers, 2009, Gazzaniga, 1995, Gazzaniga, 2000, Gopnik, 1993, Wegner, 2002, Wilson, 2002]. This view makes several testable predictions, for example, that people with a good mentalizing ability should also have good metacognitive ability; and that if children have problems with inferring the mental states of others (e.g., because of a neurodevelopmental condition such as autism), they may also develop difficulties with understanding their own minds.

The second prediction can be directly studied in the context of Autism Spectrum Condition (ASC)—a neurodevelopmental condition that is, in part, characterised by nonverbal and verbal communicative problems, untypical socio-emotional reciprocity [American Psychiatric Association, 2013] and mentalizing difficulties [Baron-Cohen et al., 1985]. If the mentalizing-is-prior view is correct, difficulties with understanding other people's thoughts and social communication (as is typical in autism), should also affect the development of metacognition in this condition.

Metacognition is often quantified in laboratory tasks as the ability to provide accurate confidence ratings about self-performance in a range of cognitive domains. “Good” metacognitive ability is indicated by reporting lower confidence when wrong, and higher confidence when right [Fleming et al., 2010, Fleming and Lau, 2014, Frith and Frith, 2012, Yeung and Summerfield, 2012]. This is known as metacognitive ”sensitivity” and is distinct from metacognitive ”bias”, the tendency to be more or less confident overall [Fleming and Lau, 2014]. Mentalizing, on the other hand, is often assessed as participants’ ability to understand what agents are thinking or intending from observations of their expressions [Abell et al., 2000, Baron-Cohen et al., 2001, White et al., 2011]. ”Good” mentalizing ability is indicated by correct assessment of others’ mental states. To date, six studies have examined associations between metacognition and mentalizing in children or adults with autism [Carpenter et al., 2019b, Grainger et al., 2016a, Nicholson et al., 2019, Nicholson et al., 2020, Wojcik et al., 2013, Williams et al., 2018]. Three of the six papers suggest, in line with the idea that mentalizing and metacognition have a similar neurocomputational mechanism, that autistic individuals have metacognitive difficulties that are commensurate with their mentalizing capacity [Grainger et al., 2016a, Nicholson et al., 2020, Williams et al., 2018]. However, the remaining thee studies did not find untypical metacognition in autistic compared with non-autistic participants, despite finding poorer mentalizing efficiency [Wojcik et al., 2013, Carpenter et al., 2019b]. Taken together, the existing data indicate a link between metacognition and mentalizing, but not equivocally so.

One difficulty with interpreting findings on metacognition is that its measurement is often confounded by other aspects of task performance, which itself may vary across individuals and clinical groups. For example, many of the studies reviewed above computed people’s metacognitive sensitivity as the Goodman-Kruskall gamma correlation between trial-by-trial accuracy and confidence [Nelson, 1984], a measure known to be confounded by *type 1 sensitivity* (task performance) and *metacognitive bias* (people’s average confidence scores; [Fleming and Lau, 2014, Maniscalco and Lau, 2012, Maniscalco and Lau, 2014, Masson and Rotello, 2009, Rahnev and Fleming, 2019]; **Figure 2.1a**). The impact of this confound may be particularly pertinent in studies comparing autistic and non-autistic people, as sensory (hyper-) sensitivity [Ewbank et al., 2016, Lieder et al., 2019, Pirrone et al., 2017] and over-confidence [McMahon et al., 2016, Milne et al., 2002, Zalla et al., 2015] are sometimes found to be higher in autistic compared to non-autistic groups. In other words, previously reported measures of *metacognitive* sensitivity may have been confounded by higher *sensory* sensitivity in autistic participants.

A powerful approach to control for task performance confounds in studies of metacognition is to use model-based metrics derived from signal detection theory, that allow metacognitive sensitivity to be expressed in the same units as task performance, while also controlling for metacognitive bias (meta-*d'*; [Maniscalco and Lau, 2012, Maniscalco and Lau, 2014]). Notably, a recent study identifying a positive correlation between metacognitive and mentalizing efficiency when using this meta-*d'* metric to quantify metacognitive sensitivity [Nicholson et al., 2020]. Nicholson and colleagues (2020) measured both implicit (behavioural) and explicit (verbal) metrics of choice uncertainty (defined as 'opting out' from choosing or verbally reporting lower confidence, respectively) and measured mentalizing efficiency from participants' descriptions of short animations of abstract figures that vary in their level of intentionality [Abell et al., 2000]. The authors found that explicit, but not implicit, metacognitive sensitivity was positively correlated with mentalizing efficiency, and significantly lower among autistic children. In a second study on neurotypical adults, the authors leveraged a dual-task condition in which participants completed a mentalizing or non-mentalizing-related cognitive task alongside a metacognition task and found that the dual mentalizing task significantly lowered metacognitive sensitivity compared to conditions in which the dual task did not require mentalizing [Nicholson et al., 2020]. Together these findings suggested that mentalizing and metacognitive ability share a common neurocognitive basis which is commensurately impaired in autistic individuals.

However, despite this promising result, further limitations in the measurement of both mentalizing and metacognition in Nicholson et al (2020) are worth considering. First, mentalizing efficiency was scored from participants' written descriptions of the triangles' mental states. It has been proposed that this type of question is more prone to confounds of verbal fluency than, for example, multiple-choice assessments of mentalizing [White et al., 2011]. This may be particularly problematic in studies of autism given that differences in verbal fluency are commonly observed in this condition [Livingston et al., 2019b, Spek et al., 2009]. Second, in the metacognition task, decisions were of varying choice difficulty, with some perceptual discriminations (of colour, or dot density) being easier than others. When task difficulty is varying between trials and participants, it may affect measures of metacognitive ability, even when d' is controlled for [Rahnev and Fleming, 2019]. Finally, participants received trial-by-trial feedback on their confidence ratings, where they were rewarded for reporting higher confidence on correct trials correct trials and lower confidence on error trials (i.e., better metacognition was incentivized). This may have created a disadvantage for autistic participants who may have difficulties with interpreting and learning from ambiguous or implicit feedback [Broadbent and Stokes, 2013, Greene et al., 2019, Reed, 2019, Robic et al., 2015, Sapey-Triomphe et al., 2018, Zwart et al., 2018]. In other words, it could be that the lower metacognitive ability in the autistic group was a consequence of failing to maximize rewards on the basis of the ambiguous feedback.

In this study, I set out to control for some of the factors that might have influenced the results of these previous studies by adopting experimental and computational methods that are considered optimal for the assessment of metacognitive sensitivity [Fleming, 2017, Rahnev and Fleming, 2019]. Specifically, I measured metacognition using a psychophysical task on which participants make repeated perceptual judgements and rated their confidence in being correct. In order to match sensory sensitivity across participants and over the course of the experiment within the same participant, I employed a staircase procedure that continually adjusted sensory evidence strength on the basis of people's responses. In addition, I measured these same participants' mentalizing efficiency on a separate task in which they watched short animations of abstract figures that moved across the screen according to distinct types of interaction [Abell et al., 2000], similar to that used by Nicholson et al (2020; **Figure 2.1b**). Instead of providing a verbal description of each interaction, participants indicated their answer using multiple choice selection [White et al., 2011]. I controlled for type 1 performance in the measurement of metacognition by computing *metacognitive efficiency* (meta-*d'/d'*), which controls for type 1 sensitivity and metacognitive bias using the meta-*d'* model [Maniscalco and Lau, 2012, Maniscalco and Lau, 2014]. Moreover, I estimated metacognitive efficiency within a Bayesian hierarchical model that allows optimal estimation of the relationship between metacognitive efficiency and individual differences in mentalizing efficiency, while also taking into account uncertainty surrounding each individual subject's parameter estimates [Fleming, 2017, Harrison et al., 2020].

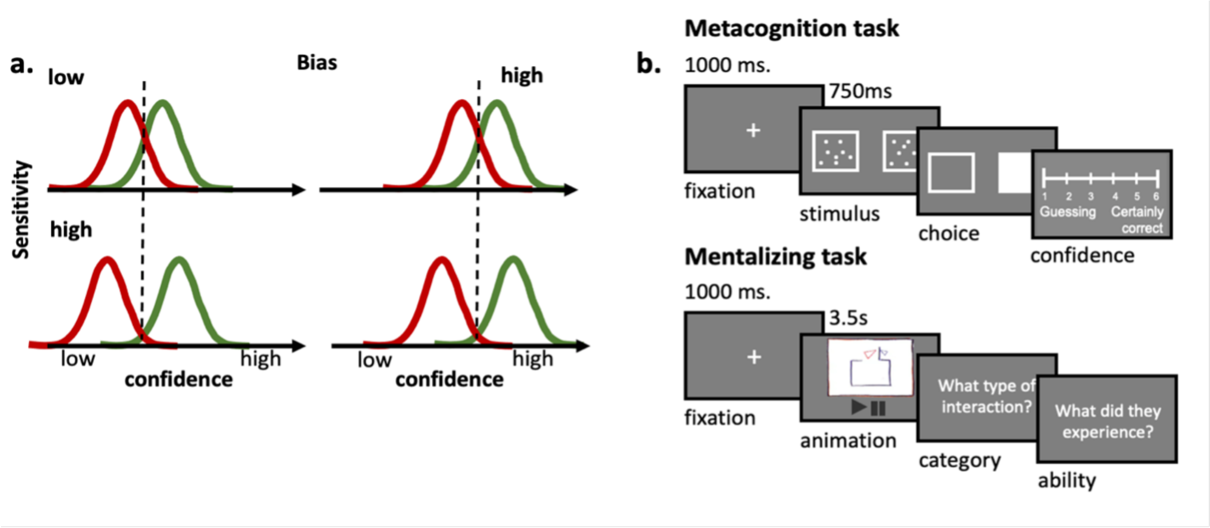


Figure 2.1: ***Task design and dissociation between metacognitive sensitivity and bias. (a)*** *Hypothetical Gaussian distributions of confidence for correct (green) and incorrect (red) decisions. The left panel represents a decider with low confidence; the right panel represents a decider with high confidence. Metacognitive sensitivity is the separation in confidence between correct and incorrect decisions; metacognitive bias is the overall confidence expressed.* ***(b)*** *On the metacognition task, participants made judgments about which patch with dots had a higher density (left or right). After this, they were asked to rate their confidence on a scale from 1 "Guessing" to 6 "Certainly correct". On the mentalizing task, participants watched animations of moving triangles and were asked to categorize and interpret the interaction of the triangles.*

In a second set of analyses, I investigate *how* the computation of confidence is modulated by mentalizing efficiency by building hierarchical regression models of trial-by-trial confidence ratings. I reasoned that, if metacognition and mentalizing rely on the same inferential processes and cues, mentalizing efficiency should facilitate the use of behavioural cues that are similarly predictive of the mental states of others. Work in cognitive psychology has shown that response times have a causal impact on the confidence levels people ascribe not only to themselves, but also to others [Patel et al., 2012, Palser et al., 2018]. In a series of exploratory analyses, I therefore asked whether confidence was more tightly coupled to response times among participants with better mentalizing efficiency.

Following the pre-registered methods and hypotheses, I recruited N=501 adults via the online research platform Prolific (www.prolific.co). To pre-empt my results, I found that (1) metacognitive and mentalizing efficiency are positively correlated, even after controlling for first-order performance; (2) mentalizing efficiency is associated with greater coupling between response time and confidence, suggesting those with greater mentalizing are more sensitive to inferential cues to confidence, and (3) metacognitive efficiency is lower in those with having greater difficulties with social communication and understanding. Taken together, these results suggest that the ability to reflect on our own mental states is modulated by one's ability to access and understand the thoughts and mental processes of others.

2.2 Methods

2.2.1 Participants

I recruited N=501 proficient English speaking participants via Prolific, a recruitment platform more representative of real populations than standard student samples [Palan and Schitter, 2018]. All participants accessed the experiment with a desktop computer or laptop (no tablets or smartphones). Exclusion criteria were responding incorrectly a "catch" question (e.g., "If you are still paying attention, please select x as your answer"); performing below or above pre-defined accuracy cut-offs (60% and 90% respectively); and rating the same confidence on more than 90% of metacognition trials. This resulted in the exclusion of N=23 participants (5%), leaving N=477 participants for further analysis (198 female, mean age: 28.73 ± 0.52 years). All participants gave informed consent before the experiment, which was approved by the University College London Ethics Committee (1260/003).

2.2.2 Metacognition Task

Stimuli were programmed in JavaScript using JSPsych (version 5.0.3) and hosted on the online research platform Gorilla (https://gorilla.sc/). Participants made 168 decisions across four blocks concerning which box was filled with a higher density of dots (left or right, indicated by pressing the "W" or "E" key, respectively without a time limit). The boxes were two black squares (each 250 x 250 pixels) which were each subdivided into grids of 625 cells that were filled with 313 dots. Choice difficulty was manipulated by adjusting the dot difference between boxes according to a 2-down-1-up staircase procedure: dot difference increased after every error and decreased after two consecutive correct answers. Dots seemed to flicker, an effect created by replotting five different configurations of the same dot difference level for 150 ms each, for a full stimulus duration of 750 ms [Rollwage et al., 2018]. On 26 practice trials participants received immediate feedback. During the remaining trials, participants did not receive feedback but had to rate their confidence that their decision was correct (on a scale from 1 "*Guessing*" to 6 "*Certainly Correct*", without a time limit; [Rouault et al., 2018]). The total duration of the metacognition task was approximately 20-30 minutes and participants were instructed to take six self-paced breaks at specific moments in the task.

2.2.3 Mentalizing Task

I used a validated online version of the Frith-Happé Triangle Task [Abell et al., 2000]. Participants were shown twelve short (34-35 sec.) animations of one large red and one small blue triangle. The way in which the triangles moved was manipulated across three conditions: in random animations they moved purposelessly around; in Goal-Directed animations they interacted behaviourally; and in four Theory of Mind (ToM) animations they interacted in a way that involves responding to the other's mental states. Participants were scored on their accuracy in classifying which category the interaction pertained to (mentalizing classification) giving a score ranging between 0-12 (i.e., participants could score one point after each animation). In addition, I computed participants' accuracy in categorizing the feelings of the triangles (mentalizing efficiency; [White et al., 2011]). Mentalizing efficiency was scored as the proportion of correctly identified mental states of the triangles, given that the ToM animation had been correctly identified. Specifically, I divided the number of correctly identified mental states within the ToM animations by the number of correctly identified ToM animations. This metric contains more variation than just the proportion of correctly classified ToM animations alone, as it measures the ability to both identify and understand mental states [White et al., 2011]. Furthermore, this type of mental state attribution requires tracking the triangle's intentions throughout the animation and cannot simply be deduced from the general kinematics of the triangle, therefore making it less susceptible to compensatory strategies. Participants had to watch the complete animation before the questions appeared, after which they were allowed to decide without a time limit. All animations were presented in pseudo-randomized order and after three practice animations on which participants received immediate feedback. The total duration of the task was approximately 10-20 minutes without breaks.

2.2.4 Additional measures

After the two computer tasks, which were presented in counterbalanced order, the following questionnaires were administered: (1) the Autism Quotient-10 (AQ10), a brief assessment of autistic traits (where a higher score means more autistic traits; [Allison et al., 2012]); (2) the RAADS-14, a screening tool for autistic traits in adult populations which asks whether each trait was present either in childhood, currently, both or neither (where a higher score means more autistic traits; [Eriksson, 2013]), (3) the Beck Cognitive Insight Scale (BCIS), an assessment of people's ability to distinguish objective reality from subjective experience [Beck et al., 2004] and (4) the International Cognitive Ability Resource (ICAR), a brief assessment of fluid intelligence [Condon and Revelle, 2016].

2.2.5 Statistics

The hypotheses and analyses for this study were pre-registered (https://osf.io/vgy7a/). Validation checks consisted of Spearman's rho correlations (which are recommended for ordinal data) to assess relationships between main composite survey scores. Equal variances were assumed if not otherwise specified. I report *P* values at a 0.05 alpha level and the 95% confidence interval (95% CI) of the test statistic. Type-1 cognitive and type-2 metacognitive parameters were estimated using the open source HMeta-d toolbox (https://github.com/metacoglab/Hmeta-d) implemented in MATLAB (version 9.7.0). Type-2 *meta-d'*, the ability to determine one's accuracy with confidence ratings, was inferred using Markov chain Monte Carlo (MCMC) Bayesian sampling procedures using JAGS (http://mcmc-jags.sourceforge.net) across 30,000 samples after a burn-in of 1,000 samples distributed across three chains. My parameter of interest was $Mratio$ *meta-d/d*, or metacognitive efficiency, which expresses metacognitive sensitivity (*meta-d'*) relative to task performance (*d'*, in other words, an Mratio of 1 implies participants have optimal metacognitive efficiency; [Fleming, 2017]).

I assessed model convergence for each HMeta-d model by ensuring that the consistency of the posteriors within and between chains, the Gelman-Rubin (G-R) statistic, was below 1.1 [Andrew Gelman and Donald B. Rubin, 1992] and by visually inspecting the chains (**Figure 2.2a**). In addition, each reported model was checked for reliability by conducting posterior predictive checks, the extent to which the model parameters could recover key patterns of the data (**Figure 2.2b**).

To test the first pre-registered hypothesis of a positive association between metacognitive and mentalizing efficiency, I incorporated a simultaneous hierarchical estimation of the beta coefficient (β) of the impact of my standardized mentalizing efficiency score, $Menta$, on the log of metacognitive efficiency, *log(Mratio)*. A log-transform of Mratio was used in these analyses to allow equal weight to be given to increases and decreases from the optimal ratio of 1 [Fleming, 2017].

$$log(Mratio)_s \sim \beta_0 + \beta_1 menta_s + \varepsilon_s \tag{2.1}$$

$log(Mratio)_s$ denotes baseline group-level metacognitive efficiency, $menta_s$ is the mentalizing score for subject *s*; and ε_s refers to noise that is drawn from a T-distribution with variance σ_δ and 5 degrees of freedom, multiplied by a noise parameter ζ. I used priors found to provide the most efficient regression parameter recovery [Harrison et al., 2020] which were drawn from Gaussians $N(\mu,\sigma)$, half-Gaussians $HN(\mu,\sigma)$ or T-distributions $T(\mu,\sigma,df)$:

$$Mratio \sim N(0,1)$$

$$\beta \sim N(0,1)$$

$$\sigma_\delta \sim HN(1)$$

$$\zeta \sim \beta(1,1)$$

$$\delta_s = T(0,\sigma_\delta,5)$$

$$\varepsilon_s = \zeta * \sigma_\delta$$

The highest density interval (HDI) represents the 'credible' posterior range within which 95% of the estimated regression coefficient falls. I plotted the HDI for the regression coefficient and assessed significance by computing the probability that it differed from zero, $P_\theta(HDI < 0|HDI > 0)$, where a higher probability suggests a stronger effect [Kruschke, 2010].

I also calculate $log(Mratio)_s$ at the individual level for use in post-fit frequentist analyses. These were conducted using linear models with $log(Mratio)_s$ as the dependent variable and $menta_s$ and covariates (standardized age, IQ, gender [-1: female, 1: male] and education (edu) [1: no education, 2: high school or equivalent, 3: some college, 4: BSc, 5: MSc, 6: doctoral]):

$$\begin{gathered} log(Mratio)_s \sim \beta_0 + \beta_1 menta_s + \beta_2 age_s + \\ \beta_3 IQ_s + \beta_4 gender_s + \beta_5 edu_s + \varepsilon_s \end{gathered} \tag{2.2}$$

To test the effect of autistic traits on $log(Mratio)_s$ I ran the same models specified in Equations 2.1 and 2.2 but now replacing menta$_s$ with the $RAADS14_s$ main composite autistic trait scores [Eriksson, 2013]. In preliminary analyses I failed to replicate previous findings of a negative correlation between mentalizing efficiency and AQ-10 scores [Allison et al., 2012], and therefore (deviating from our pre-registration plan) I decided to conduct all further analsyis of questionnaire data using RAADS-14 scores alone [Bertrams, 2021].

To assess the effects of trial-by-trial standardized (log) response *logRT* and accuracy on confidence, I conducted hierarchical mixed-effect regression models using the 'lme4' package in R (version 3.3.3) and plotted the standardized fixed-effect beta coefficients of the model fits. I obtained the *P*-values of the regression coefficients using the *car* package. All models include a random effect at the participant level and all statistics are computed at the group level. I report type III Wald chi-square tests (χ^2), degrees of freedom (df) for fixed effects, and estimated beta-coefficients (β) together with their standard errors of the mean (± SEM) and *P*-values of the associated contrasts.

To investigate if *logRT* informs confidence differently as a function of individual differences in autistic traits, I test whether a hierarchical mixed-effect regression model better predicts trial-by-trial confidence (conf) when the predictor variables accuracy (acc) [-1: error, 1: correct], z-score of the log response time (RT) and their interactions (Equation 2.3) were allowed to vary as a function of individual differences in standardized autistic trait scores (*RAADS* in Equation 2.4):

$$\begin{gathered} conf \sim acc + logRT + acc * logRT + \\ (1 + acc + logRT + acc * logRT | subj) \end{gathered} \tag{2.3}$$

$$\begin{gathered} conf \sim RAADS_s * (acc + logRT + acc * logRT) + \\ (1 + acc + logRT + acc * logRT | subj) \end{gathered} \tag{2.4}$$

The results of the Likelihood Ratio Test are expressed in terms of the *Akaike Information Criterion* (*AIC*): $\Delta AIC = AIC_{\text{eq2.3}}$–$AIC_{\text{eq2.4}}$, and the *Log Likelihood*

(*LL*): $\Delta LL = LL_{\text{eq2.3}}–LL_{\text{eq2.4}}$ with associated *P* values extracted from type III Wald chi-square tests (χ^2).

For Structural Equation Modelling I use the “lavaan” package in R (version 1.2.5033) and report the standardized factor loadings (β ± SEM) with associated *P* values extracted from a type III Wald chi-square tests (χ^2).

2.3 Results

2.3.1 Performance checks and auto-correlations

The staircase converged to a stable performance level within and between participants (choice accuracy: M = 75%, SEM = 0.23, **Figure 2.2a**). Average metacognitive efficiency or Mratio (M=0.693 ± 0.016; **Figure 2.2b**) and the log of response times (logRT; M = -1.405e-17 ± 0.046; **Figure 2.2c**) were also similar to those of previous datasets [Rouault et al., 2018, Rollwage et al., 2018].

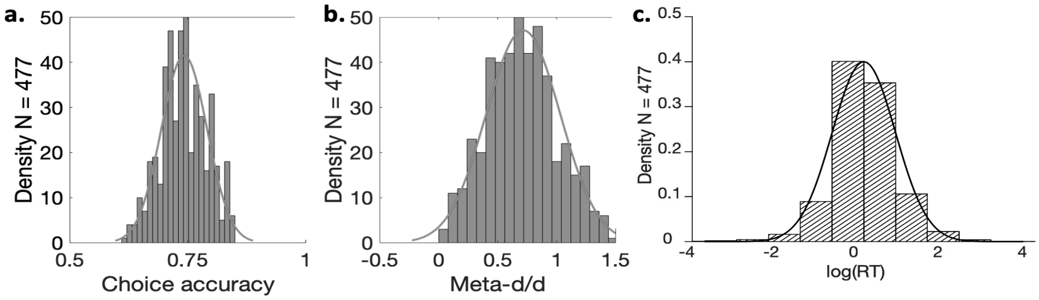


Figure 2.2: ***Choice accuracy on the metacognition task. a.*** *Histogram distribution of choice accuracy.* ***b.*** *Histogram distribution of metacognitive efficiency (meta-d'/d').* ***c.*** *Histogram distribution of the log of standardized response times (logRT). All variables are derived from the metacognition task and plotted for the group as a whole (N=477).*

Given that stimulus sensitivity can trivially affect estimates of metacognitive sensitivity [Rahnev and Fleming, 2019], I also sought to ensure key variables related to metacognition and mentalizing were independent of first-order task performance. I first calculated whether each individual’s experienced stimulus variability (the ratio between the standard deviation of stimulus difficulty and average stimulus difficulty) and correlated this with the main variable of interest. Staircase variability was not correlated with mentalizing efficiency (rs_{475} = 0.005, P = 0.91, **Figure 2.3a**); metacognitive efficiency (rs_{475} = -0.068, P = 0.137; **Figure 2.3b**); RAADS-14 scores (rs_{475} = 0.0015, P = 0.974; **Figure 2.3c**) or AQ-10 scores (rs_{475} = -0.066, P = 0.149; **Figure 2.3d**). These same validation checks were conducted for perceptual sensitivity, which was not correlated with mentalizing efficiency (rs_{475} = 0.0656, P = 0.1524; **Figure 2.3e**), metacognitive efficiency (rs_{475} = -0.0513, P = 0.264; **Figure 2.3f**), RAADS-14 scores (rs_{475} = -0.0536, P = 0.2437; **Figure 2.3g**) or AQ-10 scores (rs_{475} = 0.0359, P = 0.435;**Figure 2.3h**).

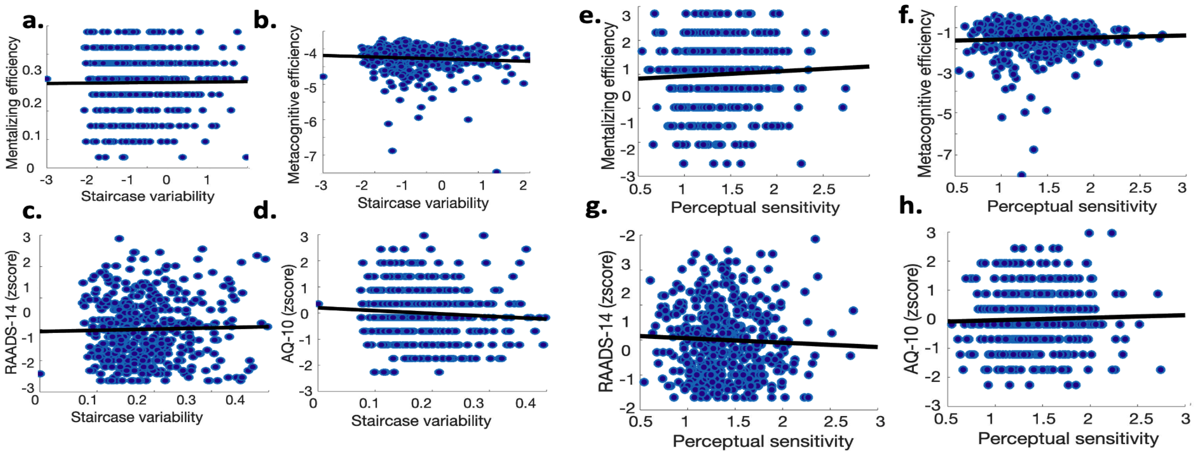


Figure 2.3: ***Correlations between the main variables of interest. a-d:*** *Staircase variability, the ratio of the standard deviation and the mean dot difference, was not correlated with:* ***a.,*** *mentalizing efficiency,* ***b.,*** *Metacognitive efficiency (meta-d'/d'),* ***c.,*** *autistic traits as measured with the RAADS-14;* ***d.,*** *autistic traits as measured with the AQ-10.* ***e-h:*** *Perceptual sensitivity (d') was not correlated with:* ***e.,*** *mentalizing efficiency,* ***f.,*** *Metacognitive efficiency (meta-d'/d'),* ***g.,*** *autistic traits as measured with the RAADS-14,* ***h.,*** *autistic traits as measured with the AQ-10.*

2.3.2 Posterior predictive checks

Next, I test whether the HMeta-d models used in estimating metacognitive efficiency were reliable by means of convergence checks and posterior predictive checks. The hierarchical regression model predicting metacognition from mentalizing efficiency scores converged well, indicated by the Gelman-Rubic statistics ($\hat{R}_{Mratio}$=0.99997 and plotted chains in **Figure 2.4a**). In addition, posterior predictive plots captured key patterns of the participants' confidence responses, with model and predicted type ROCs closely overlapping **Figure 2.4b**).

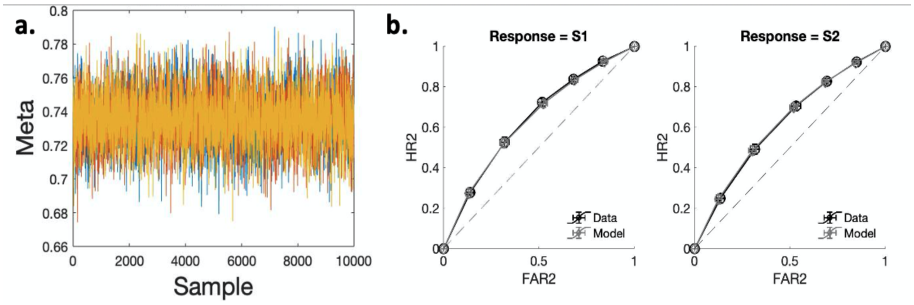


Figure 2.4: ***Posterior predictive checks on HMeta-d fits. a.*** *MCMC chains for parameter meta-d'/d' (metacognitive efficiency) from the hierarchical regression model.* ***b.*** *Observed and model estimates for the Type 2 ROC curves for leftward (S1) and rightward (S2) responses from the regression meta-d model fits. Error bars represent the mean ± standard error of the mean.*

2.3.3 Mentalizing efficiency checks

As an indication of the reliability of mentalizing task variables, I ask whether the two mentalizing measures from the Happé-Frith Triangle Task were positively correlated. This was the case, with a positive correlation between the mentalizing feelings and mentalizing category scores (Spearman's r = 0.37, P = 2.73e-16). In addition, to establish whether the autistic trait surveys and Frith-Happé triangle task were measuring a similar mentalizing construct, I tested whether people with more autistic traits on the mentalizing subscale of the RAADS-14 also had lower mentalizing efficiency on the Frith-Happé Triangle Task, which was also the case (Spearman's r = -0.11, P = 0.017).

2.3.4 A common computation for mentalizing and metacognition

Having conducted some reliability checks I next investigated the hypothesis of a positive association between metacognitive efficiency and mentalizing efficiency within the hierarchical meta-d' model. When I examined the beta coefficient representing the impact of mentalizing efficiency on metacognitive efficiency, the

HDI was positive and did not encompass zero (hierarchical estimation: 95% HDI [0.01, 0.09], with 99% of the sampled beta values being higher than zero (P_θ ($mentalizing > 0$) = 0.99; **Figure 2.5a**) indicating a significant positive relationship.

To confirm this effect while controlling for covariates of age, gender, IQ and education, I used a linear regression model with the standardized log metacognitive efficiency from a single-subject model as a dependent variable and mentalizing efficiency and these covariates as predictor variables. This approach again revealed a positive relationship between mentalizing and metacognition ($\beta_{mentalizing} = 0.11$, SE = 0.54, $t_{476} = 2.26$, $P = 0.02$) and no effects of the covariates ($P > 0.05$), suggesting that participants who were better at inferring the mental states and interactions on the mentalizing task were also better at tracking their performance on the metacognition task.

2.3.5 Using response times to inform explicit confidence

To investigate how mentalizing was related to metacognition, I next tested the hypothesis that mentalizing is associated with a greater impact of response times on confidence. Specifically, I estimated a hierarchical mixed-effects model predicting trial-by-trial explicit confidence levels on the metacognition task from differences in standardized log response times (logRT) and accuracy [error: -0.5, correct: 0.5] (Equation 2.3), and asked whether this model provided a better fit when these predictors were allowed to vary as a function of the participants' mentalizing efficiency (Equation 2.4). A Likelihood Ratio Test indicated that this was the case ($\chi^2(4)$ = 27.59, P = 1.51e-05) which was also confirmed by several goodness-of-fit indices (log likelihood (LL): ΔLL = 13,ΔAIC = -20, ΔBIC = 17 and $\Delta Deviance$ = -28), suggesting a significant relationship between mentalizing and the computations underpinning confidence formation.

I next asked how mentalizing modulated the construction of confidence by investigating which predictor variables interacted with mentalizing efficiency. I found that participants with better mentalizing efficiency reported lower overall confidence in their own responses than participants with lower mentalizing efficiency (hierarchical linear regression, main effect of mentalizing efficiency (main effect of mentalizing efficiency: $\chi^2(1)$ = 64.08, P = 0.01, β = -0.40, SE = 0.02). In addition, participants with higher scores of mentalizing efficiency scores modulated their confidence ratings more on the basis of their response times than participants with lower scores of mentalizing efficiency (interaction effect of logRT x mentalizing efficiency: $\chi^2(1)$ = 21.92, P = 2.84e-06, β = -0.03, SE = 0.006; **Figure 2.5b**, consistent with the idea that mentalizing facilitates metacognition by facilitating self-inference on the basis of externally visible behavioural cues.

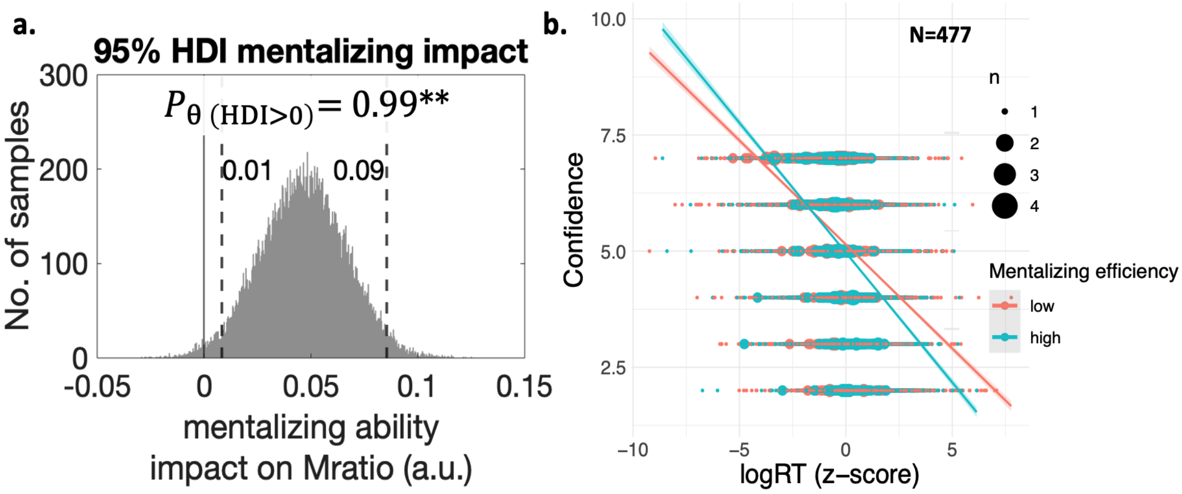


Figure 2.5: ***Mentalizing modulates computation of confidence.*** ***(a)*** *Posterior distribution over the regression coefficient relating mentalizing efficiency to metacognitive efficiency. The dashed lines represent the 95% highest density interval (HDI).* P$_\theta$ *indicates the probability that the posterior samples are greater than zero, *** P < 0.01 in the frequentist linear model.* ***(b)*** *Confidence was negatively related to response times (logRT). Trial-by-trial response times have a higher impact on the estimated confidence of participants scoring above the median of mentalizing efficiency scores (in turquoise) than participants scoring below the median (in pink). Shaded area represents the Standard Deviation from the Mean (±SDM).*

2.3.6 Exploring the relationship between metacognition and autistic traits

Next, I addressed the second hypothesis of a negative association between metacognitive efficiency and autistic traits in the general population, as assessed with the AQ-10 [Allison et al., 2012] and the RAADS-14 questionnaires [Eriksson, 2013]. First, I evaluate whether participants with higher scores of autistic traits had lower mentalizing efficiency, by conducting a linear regression model with mentalizing efficiency as the dependent variable and autistic trait scores and the covariates (age, gender, education, IQ) as predictor variables. I only found the expected negative relationship between mentalizing efficiency and RAADS-14 scores (linear regression model: $\beta_{RAADS-14}$ = -0.002, SE = 0.0009, t_{476} = -2.21, P = 0.03) but not AQ-10 scores (linear regression model: β_{AQ10} = 0.006, SE = 0.004, $t_{1.33}$, P = 0.19). This unexpected finding, together with recent re-evaluations of the reliability of the AQ-10 scale [Bertrams, 2021], and the greater developmental information captured by the RAADS-14, led me to focus on RAADS-14 scores in the remainder of the analyses.

Next, I asked whether compromised mentalizing efficiency in participants with higher scores of autistic traits was associated with lower metacognitive efficiency. To test this, I estimated the correlation between metacognitive efficiency and RAADS-14 scores within a hierarchical regression model. The 95% HDI for the coefficient of RAADS-14 scores was negative on average, ranging from [-0.057, 0.019], but encompassing zero (hierarchical estimation: P_θ ($RAADS < 0 = 0.82$). A frequentist linear model that controlled for the covariates also confirmed that participants with higher scores of autistic traits do not necessarily also have compromised metacognitive efficiency (linear regression model: $\beta_{RAADS14}$ = -0.0006, SE = 0.005, t_{476} = -1.09, P = 0.28).

An alternative explanation hypothesis is that autistic traits as measured by the RAADS-14 do not have a direct impact on the metacognitive efficiency score, but rather affect the construction of confidence. To examine this, I tested if my mixed-effect hierarchical regression model better predicts trial-by-trial confidence levels on the metacognition task when the predictors (accuracy, logRT and their interactions) were allowed to vary as a function of differences in autistic traits. A Likelihood Ratio Test indeed suggests that an interaction term on autistic traits improved the fit of the model ($\chi^2(4)$ = 14.52, P = 0.006) which was further confirmed by several goodness-of-fit metrics (Δ_{LL} = 7, Δ_{BIC} = -31, Δ_{AIC} = 7 and $\Delta_{Deviance}$ = -15), indicating that the computation of confidence differs as a function of individual differences in autistic traits.

I next asked in what way people with higher scores for autistic traits constructed their confidence differently, by testing which predictor variables interacted with RAADS-14 scores. I found that participants with higher scores for autistic traits reported lower confidence overall (hierarchical linear regression, main effect of RAADS-14: $\chi^2(1)$ = 4.86, P = 0.027, β = -0.008, SE = 0.004). In addition, explicit confidence was more informed by logRT among participants with lower scores for autistic traits than among participants with higher scores for autistic traits (interaction effect of logRT x RAADS-14: $\chi^2(1)$ = 6.46, P = 0.011, β = 0.004, SE = 0.001). In **Figure 2.6a** I plot the extracted beta coefficients of the impact of response times on confidence for participants scoring above and below the median cut-off on autistic traits on error and correct trials separately, which shows that this effect was driven by participants with higher autistic traits scores having a lower impact of response times on error-trials than participants with lower autistic traits (three-way interaction of logRT x RAADS-14 x accuracy: $\chi^2(1) = 4.63$, $P = 0.031$, β = -0.003, SE = 0.001). Together these results suggest that participants with higher autistic traits use response times less to infer they have committed an error than participants with lower autistic trait scores.

These results suggest that compromised mentalizing efficiency may specifically affect the relationship between response times and confidence. To test this, I asked if metacognitive and mentalizing efficiency were also positively correlated in the new sample. Pearson correlations revealed no significant correlation between metacognitive and mentalizing efficiency in the autistic group (Pearson's r = -0.043, P = 0.79) nor the comparison group (Pearson's r = 0.206, P = 0.20) separately, whereas I did find this positive correlation in the whole population of chapter two [N = 477]. This suggests that a sample of N = 40 is insufficiently powered to reveal a correlation between metacognitive and mentalizing efficiency. In support of this, when I took both sub-groups of the data in chapter three together, the beta coefficient representing the impact of mentalizing efficiency on metacognitive efficiency was positive and did not encompass zero (hierarchical estimation: 95% HDI [0.00, 0.196]), with 98.57% of the sampled beta values falling above zero ($P_{\theta\ (mentalizing>0)}$ = 0.986; **Figure 2.6a**) indicating a significant positive relationship. In addition, a linear regression model that controlled for the covariates confirmed a positive relationship between mentalizing and metacognition (frequentist linear regression model: $\beta_{mentalizing}$ = 0.24, SE = 0.12, t_{63} = 2.00, P = 0.049; **Figure 2.6b**) with no effects of the covariates (all $P > 0.05$). This suggests that in both autistic and non-autistic people the ability for mentalizing positively predicts the ability for metacognition.

Next, I asked whether specifically social (mentalizing and communicative) aspects of the autistic phenotype, rather than non-social aspects, negatively impact metacognition. In an exploratory analysis I estimated the correlation between metacognitive efficiency and self-reported social skills social with hierarchical regression models. This analysis revealed that participants with self-reported difficulties in everyday types of social interaction, measured by the mentalizing and social anxiety subscale of the RAADS-14 had lower metacognitive efficiency than participants with better self-reported social skills (hierarchical estimation: HDI: [-0.07, 0.00], $P_{\theta\ (socialskills<0)}$ = 0.97; frequentist linear regression: β = -0.09, SE = 0.05, t_{476} = -1.84, P = 0.067; **Figure 2.7b**). In contrast, the non-social sub-scale of the RAADS-14 (sensory reactivity) was not associated with metacognitive efficiency (hierarchical estimation: HDI: [-0.04, 0.04], $P_{\theta(non-socialskills<0)}$ = 0.43; frequen-

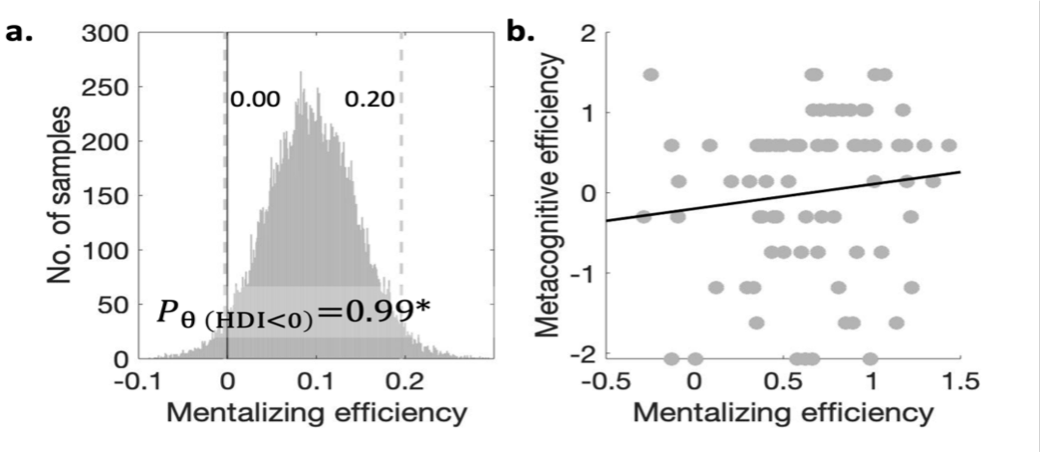


Figure 2.6: ***Mentalizing and metacognitive efficiency are positively correlated when collapsing across autistic and comparison participants.*** ***(a)*** *Posterior distribution over the regression coefficient relating mentalizing efficiency to metacognitive ability in both autistic [N = 40] and comparison [N = 40] participants. The dashed lines represent the 95% highest density interval (HDI), P_θ indicates the probability that the posterior samples are greater than zero, ** $P <$ 0.01 in the frequentist linear model.* ***(b)*** *Metacognitive efficiency is positively correlated with mentalizing efficiency when collapsing across autistic and comparison participants.*

tist linear regression: β = -0.007, SE = 0.05, t_{476} = -1.14, P = 0.89; **Figure 2.7c**). Together, these results suggest that self-reported social, but not non-social, autistic traits are negatively associated with metacognitive efficiency.

2.3.7 Structural model of a meta-representational system

One caveat of testing metacognitive and mentalizing efficiency is that their metrics are unavoidably prone to noise and measurement error. In the following analyses, I tried to circumvent this by estimating a latent variable representing one common metacognitive faculty consisting of the shared covariance between metacognitive and mentalizing efficiency and unexplained residual error. Using Structural Equation Modelling I asked whether variability in this metacognitive faculty better explains variability in social communication and understanding sub-scales than metacognitive and mentalizing efficiency separately. For these analyses, I build upon two proposed structural relationships of how access to social communication

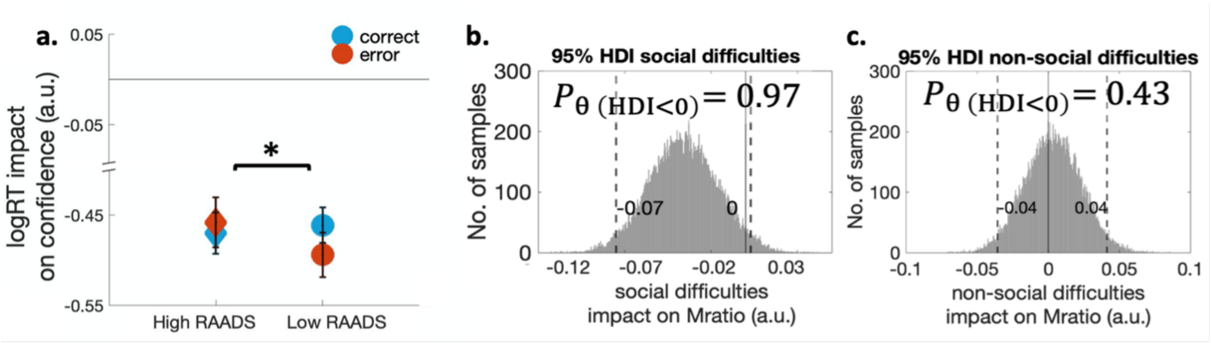


Figure 2.7: ***Autistic trait differences modulate metacognitive efficiency.*** ***(a)*** *Standardized beta coefficients of the impact of logRT on confidence from a hierarchical mixed-effect regression model on error trials (red) and correct trials (blue) for participants with high and low RAADS scores (above and below the median cut-off, respectively).* ***(b)*** *Posterior estimates of the hierarchically estimated beta coefficient relating the social subscale of RAADS-14 to metacognitive efficiency.* ***c.*** *Posterior estimates of the hierarchically estimated beta coefficient relating the non-social subscale of RAADS-14 to metacognitive efficiency. The dashed lines represent the 95% highest density intervals (HDI), P_θ indicates the probability that the posterior samples are different from zero. Error bars represent group means ± SEM, *P < 0.05 of the interaction-effect between RAADS and logRT on confidence.*

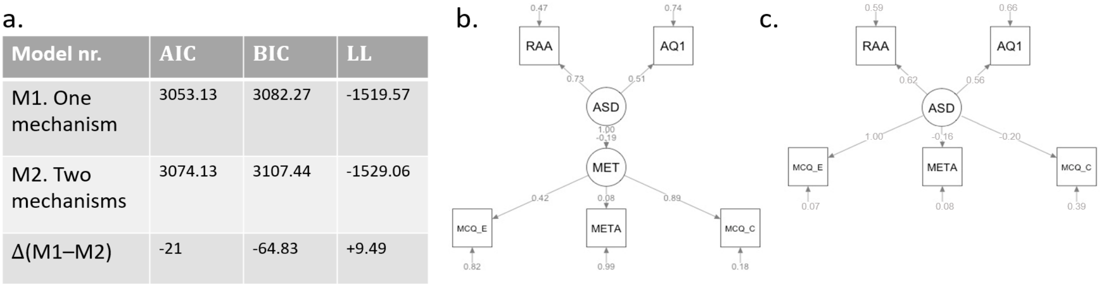
and autistic traits impact metacognitive ability [Carruthers, 2009]. The one mechanism, two modes of access account describes that there is one meta-representational system that uses social information about both what other people and oneself feel, and that this system is compromised in autism **Figure 2.8b**. The two mechanisms account describes autistic traits as two separate metacognitive and mentalizing faculties **Figure 2.8c**. Having this well-powered dataset at hand, I next wanted to explore whether I could find empirical evidence for these distinct theories.

I structurally described the one Mechanism, two modes of access model (M1) as one latent variable representing a meta-representational faculty that consists of the shared variance between mentalizing efficiency, metacognitive efficiency, and some unexplained residual error. Variance in this meta-representational faculty is a predictor (modelled as a regression) of a latent variable of autistic traits, which is described as the shared co-variance between AQ10 social communication sub-scale scores, RAADS-14 mentalizing sub-scale scores and some unexplained residual.

The Two Mechanisms model (M2), does not have a latent variable representing a meta-representational faculty. Instead, observed mentalizing efficiency and metacognitive efficiency are each individual predictors (modelled as a regression) on the same latent variable of autistic traits, which is again described as the shared co-variance between AQ10 social communication sub-scale scores, RAADS14 mentalizing sub-scale scores and some unexplained residual.

It has been advised to report an absolute index of model fit [Hu and Bentler,], to compare the fit of the model against baseline model (a model in which all observed variables are allowed to covary with all other variables). For both models, the square-root of the difference between the Root Mean Square Error of Approximation (RMSEA) are well-below the crucial cutoff score at 0.08 (M1 = 0.049, 95% CI = [0.000, 0.0947]; M2 = 0.000, 95% CI = [000, 0.08]). In addition, log likelihood scores are higher for the fitted model than for the baseline model: $\Delta LL_{M1} = 5$ and $\Delta LL_{M2} = 1$, suggesting a good model fit. To test this directly, I conducted goodness of fit tests which test whether the observed data is significantly different from data simulated on the basis of my model. This showed that the one mechanism, two model of access model provided a good fit to my data (i.e., the model simulations were not different from the observed data; $\chi^2(4) = 8.19$, $P = 0.09$, Standardized Root Mean Square Residual [SRMR] = 0.037, Comparative Fit Index [CFI] = 0.9), whereas the two mechanisms model did not provide a good fit to the data (i.e., the model simulations were statistically different from the observed data; $\chi^2(7) = 79.90$, $P < 0.001$, SRMR = 0.010, CFI = 1.00).

In line with the finding that M1 made accurate predictions of the data whereas M2 did not, model comparisons consistently showed that the M1 model fitted the data better than the M2 model (ΔLL: -1841.96,ΔBIC: 3663.28,ΔAIC: 3671.91; **Figure 2.8a**). In **Figure 2.8b** I graphically visualize the one mechanism model and its factor loadings. As expected, this model revealed a significantly negative correlation between the meta-representational faculty (where higher values represent a greater ability to take the perspectives of self and others) on a latent construct of autistic traits (β = -0.194, SEM = -.093, z = -2.07, P = 0.038). This suggests that one meta-representational faculty that receives input from both metacognition and mentalizing is negatively associated with autistic traits. In other words, access to information about other people's mental states sharpens a meta-representational faculty that is involved in both understanding one's own and other people's mental states.



a.

Model nr.	AIC	BIC	LL
M1. One mechanism	3053.13	3082.27	-1519.57
M2. Two mechanisms	3074.13	3107.44	-1529.06
Δ(M1–M2)	-21	-64.83	+9.49

Figure 2.8: ***Structural equation modelling a meta-representational faculty.*** ***(a)*** *(a) Model comparison of two proposed structural relationships between metacognition and mentalizing on autistic traits (M1 and M2). The associated goodness-of-fit estimates (AIC, BIC, LL) are indicated for each model and the difference between the models. Each of these estimates favor the M1 model.* ***(b)****. Structural overview of the one mechanism, two modes of access model, in which a latent variable representing a meta-representational faculty loads negatively onto a latent variable of autistic traits.* ***(c)****. Structural overview of the two mechanisms model, in which two latent variables, representing a metacognitive and a mentalizing faculty, separately load onto a latent variable of autistic traits. Factor loadings and ±SEM are standardized, *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$*

2.4 Discussion

In this study I used a model-based approach to show that mentalizing efficiency is positively related to metacognition. In a general population sample of N = 477 participants I found that individuals who were better at self-reported social skills and mentalizing could also more reliably track their own accuracy on a perceptual discrimination task. By investigating the trial-by-trial computations of confidence, I was able to investigate precisely how mentalizing relates to metacognition. Notably, mentalizing efficiency was associated with a tighter coupling between response times and confidence, suggesting that mentalizing efficiency may facilitate inference on cues to self-performance. These findings suggest that processes involved in inferring other people's mental states may also facilitate self-directed metacognition, and vice versa.

I quantified metacognition as the ability to reliably separate correct from incorrect decisions with confidence ratings [Flavell, 1979, Fleming et al., 2010, Rollwage et al., 2018, Rouault et al., 2018]. Several studies have suggested confidence is 'read out' from how much reliable evidence has been seen, either during the course of the decision itself [Kiani and Shadlen, 2009, Pleskac and Busemeyer, 2010] or after an initial decision has been made (post-decisional evidence processing; [Fleming et al., 2018, Resulaj et al., 2009, Talluri et al., 2018, van den Berg et al., 2016]). Other studies suggest that response times also provide a behavioural cue to confidence [Kiani et al., 2014, Patel et al., 2012]. How, then, might mentalizing play a role in confidence construction? Recent theories suggest that confidence estimates reflect an inference about the state of the decider, informed by behavioural and cognitive cues—suggesting a parallel between self- and other evaluation [Carruthers, 2009, Fleming and Daw, 2017]. Indeed, evidence strength [Pescetelli and Yeung, 2021] and response times [Patel et al., 2012] appear to be used similarly to infer both one's own and others' confidence. However, isolating such metacognitive capacity requires tight control over the evidence going into a decision, to avoid first-order performance and stimulus factors confounding estimates of the confidence-accuracy correlation [Masson and Rotello, 2009, Rahnev and Fleming, 2019]. Here, I used a staircase procedure to control perceptual performance within a narrow range and used a metric of metacognition that is unconfounded by both metacognitive bias **Figure 2.1a**) and first-order performance. In addition, I used a Bayesian inference approach to estimate the impact of mentalizing efficiency on metacognitive ability within the same hierarchical model, which ensured that both within- and between-subject variability are appropriately taken into account. These methodological advances may explain why I found a more robust between-subjects relationship between metacognition and mentalizing than reported previously [Carpenter et al., 2019b, Nicholson et al., 2020, Nicholson et al., 2019].

This work goes beyond estimating correlations between metacognition and mentalizing by revealing a potential mechanism through which mentalizing may affect metacognitive processes. Specifically, I show that better mentalizing efficiency is associated with a tighter coupling between response times and confidence. Previous work has experimentally manipulated response times and found this to have a causal effect on the construction of confidence: when response times are manipulated to be faster, people are subsequently more likely to report being confident [Kiani et al., 2014, Palser et al., 2018]. The mentalizing-is-prior theory suggests metacognition consists of a re-application of inferential processes used to understand other people to understand our own mental states [Carruthers, 2009]. These findings are consistent with this view, showing that people with greater proficiency in self-reported social skills and objectively measured mentalizing also had better metacognitive efficiency. In addition, I found that mentalizing efficiency not only correlated with overall metacognitive efficiency, but specifically with the ability to infer confidence from behavioural cues that would also be visible markers of other people's decision confidence in everyday situations. An important limitation of the current study is that I cannot draw causal conclusions about how mentalizing affects metacognition or vice-versa. Future longitudinal work is needed to ask whether exposure to situations requiring mental state inference from behaviour causally affect the development of explicit metacognition. For example, one possibility is that these effects are driven by a third factor that drives both faster response times as well higher confidence ratings, such as choice cautiousness [Balsdon et al., 2020, Khalvati et al., 2021].

Albeit associations between metacognition or mentalizing and autistic traits have been reported in isolation before, I here also explored the structural relationship between both metrics. Inspired by work suggesting a meta-representational sharpening as a function of exposure to communication about mental states [Carruthers, 2009], I directly compared how well two commonly proposed structural models of mentalizing and metacognition fitted my data. The "One Mechanism" model (M1) proposes that there is one meta-representational faculty with distinct information channels (originating from the self in the case of metacognition, or from others in the case of mentalizing) and a unique association with socio-communicative spectrum phenotypes [Frith and Happé, 1999, Happé, 2003]. On the other hand, the "Two Mechanisms" model (M2) proposes that metacognition and mentalizing are two independent capacities, which each use distinct types of information and have distinct behavioral effects [Nichols and Stich, 2003]. Using Structural Equation Modelling I was able to directly test whether, and how well, these models could be supported empirically. I found that a model that represented metacognition and mentalizing as a common meta-representational faculty most parsimoniously explained the data. In this model, a metacognitive faculty ("MET" in **Figure 2.8b**) negatively loaded onto a latent variable representing a socio-communication spectrum phenotype ("ASD" in **Figure 2.8b**), suggesting that exposure to expressions of thought and feelings may selectively strengthen a common, meta-representational system. Strikingly, the metacognitive faculty itself was informed by both the ability to identify ("MCQ C" in **Figure 2.8b** and understand ("MCQ E" in **Figure 2.8b**) the mental states of others on the mentalizing task to a greater extent than the ability to monitor one's own mental states on the metacognition task ("META" in Figure **Figure 2.8b**). This suggests, in line with the mentalizing-is-prior view, that it is primarily an understanding of other people's mental states that drives this similar understanding in oneself. Because SEM cannot make causal claims these results do not differentially favour the one mechanism, two modes of access and the mentalizing-is-prior view. However, they do suggest that these theories are more plausible than the two mechanisms theory, as the shared covariance between mentalizing and metacognitive efficiency better predicts autistic traits than both constructs entered separately. Given that making causal inferences is not recommended with structural equation modelling [Bollen and Pearl, 2013] I could not directly compare the mentalizing-is-prior and metacognition-is-prior hypotheses [Carruthers, 2009]. Longitudinal studies that track the development of both metacognition and mentalizing in children may be better suited for this type of question.

The Frith-Happé triangle task has been extensively used in autistic populations, both in controlled experimental settings [White et al., 2011, Wilson and Bishop, 2020, Abell et al., 2000] as well as in online studies of the multiple-choice version that was used in this experiment [Livingston et al., 2021]. The original task has good test-retest reliability [Brewer et al., 2017]. However, such reliability analyses have, to my knowledge, not yet been conducted on the online multiple-choice version. Future work could address this, either by testing the same participants at different time points, or by collecting different metrics of mentalizing efficiency [Nicholson et al., 2020, Nicholson et al., 2019, Grainger et al., 2016b], such as those afforded by game theoretic experimental designs [Hampton, 2008].

In summary, I here found a metacognitive benefit for participants with better mentalizing efficiency. I further disentangled the mechanism of this effect by showing that mentalizing efficiency is associated with a tighter coupling between response times and confidence in errors. Participants with better social skills were also better at reflecting upon their own performance. Together, these results suggest that inferring other people's mental states is related to the ability to evaluate our own decisions. In contrast with my pre-registered expectations, I did not find a negative correlation between metacognitive efficiency and autistic traits. This could be caused by autistic traits in the general population not being pronounced enough to engender differences in metacognitive efficiency. In the next chapter, I try to circumvent this issue by testing a clinically diagnosed autistic group and an IQ, gender, education and age-matched comparison group.

Chapter 3

Autism as a metacognitive condition

3.1 Introduction

Autism spectrum condition (ASC) is a neurodevelopmental condition that is, in part, characterised by social communication difficulties, repetitive behaviours and/or restricted interests [American Psychological Association, 2013]. Autistic people (or 'people on the autism spectrum'; [Kenny et al., 2016]) were coined to suffer from general 'mindblindness' in 1985 [Baron-Cohen et al., 1985, Baron-Cohen et al., 1986, Happé, 1994, Happé, 2003], but, since then, only a handful of studies have extended the study of mentalizing in autism to that of metacognitive efficiency about one's own behaviour and mental states [Carpenter et al., 2019b, Grainger et al., 2016a, Nicholson et al., 2019, Nicholson et al., 2020, Williams et al., 2018, Wojcik et al., 2013]. Some of these studies [Grainger et al., 2016a, Nicholson et al., 2020, Williams et al., 2018] but not others [Carpenter et al., 2019b, Wojcik et al., 2013], found that mentalizing and metacognitive efficiency were commensurately compromised in ASC.

In the previous chapter, I found that metacognitive and mentalizing abilities are related, potentially by affecting the extent to which response times modulate confidence. Against my prediction, I did not find a statistically significant negative correlation between autistic traits and metacognitive efficiency. One explanation for the null result is that the variation in autistic traits was not pronounced enough in my general population sample to allow estimation of the relationship. Another explanation is that metacognitive efficiency in autism is may not be worse on average but rather more extreme (both extremely strong and weak; [Pariser, 1981, Shields-Wolfe and Gallagher, 1992])—leading to greater variance in estimates of metacognitive efficiency in participants with more autistic traits. It could be that some participants with higher scores of autistic traits still achieved a reliable level of confidence, albeit by relying on alternative, perhaps more cognitively demanding, processes to compensate for metacognitive difficulties [Livingston et al., 2019a, Livingston et al., 2019b]. This may explain why, despite finding no association between autistic traits and metacognitive efficiency, I consistently found participants that scored higher on autistic traits to engage in distinct constructions of confidence than comparison participants. A final explanation is that metacognitive efficiency is not associated with the complete spectrum of autistic traits, but only with those traits that affect communicative and social abilities in autism. Autism encompasses both social and non-social symptoms, which are genetically [Robinson et al., 2016, St Pourcain et al., 2018, St Pourcain et al., 2014, Warrier et al., 2019] and behaviorally [Howlin et al., 2004, Howlin et al., 2013, Moss et al., 2017] dissociable. It could be that metacognitive problems in autism are restricted to individuals that have selective difficulties with social, but not non-social, autistic phenotypes.

In this chapter, I investigate whether metacognitive efficiency differs between autistic and typically developing participants and whether mentalizing and metacognitive abilities is commensurately compromised in autistic people. Following my pre-registered methods and hypotheses, I recruit forty participants between 18 and 50 years old with a self-reported diagnosis of ASC by a health care professional, on the same experimental task that was used in chapter two. To pre-empt my results, when I compare a group of diagnosed autistic and comparison participants, I indeed found that metacognitive efficiency was lower in the autistic group than in the comparison group. In addition, I replicated the result of chapter two that autistic participants relied less on their response times to infer confidence on error trials than on correct trials, indicating that mentalizing difficulties in this condition may generalize to difficulties with metacognition.

3.2 Methods

3.2.1 Participants

This study was approved by the Research Ethics Office of King's College London [HR-19/20-17704]. I recruited N = 43 autistic participants via the research charity *Autistica* (www.autistica.org.uk). Interested participants first completed an online pre-screening questionnaire that included questions about mental health and demographics. Participants that met the inclusion criteria (i.e., aged between 18 and 50 years old and a self-reported disorder by a health professional) were sent a link to the online experiment that could be accessed with a desktop computer or laptop (no tablets or smartphones). Exclusion criteria were the same as in chapter two. Three participants were excluded: one participant performed below the a priori accuracy cut-off and two participants performed above the a priori accuracy cut-off. This resulted in the exclusion of N = 3 participants (7.5% of the total sample, which is consistent with chapter two), leaving data from N=40 participants for analysis (37 female, mean age: 37.90, SEM = 1.59 years). All participants gave informed consent before experiment onset.

To obtain an equal number of comparison participants I re-analysed the dataset of chapter two that included the same task and questionnaire battery. This dataset consisted of N = 477 English-speaking participants from the general population (198 female, mean age: 28.73 ± 0.52 years). Data on mental health conditions was not collected. To ensure that the participants selected from this former dataset provided a comparison group with low autistic traits, I first reduced the dataset to N = 97 participants scoring in the lowest 50% of the RAADS-14 and AQ-10 responses (a score lower than 16 and 5, respectively, which is more stringent than the recommended clinical cut-off score; [Ashwood et al., 2016, Eriksson et al., 2013]). Next, to ensure the groups were well-matched on other characteristics, for each included autistic participant I manually selected a comparison participant of similar gender (a high proportion of females in the autism group meant that it was not possible to find a 1:1 gender match for 3 participants); who was within ±5 years from the target age; ±2 levels from the target education; and ±5 ICAR points from the target fluid intelligence level. These criteria were identified after initial exploration indicated they provided sufficient flexibility to provide a reasonable match between the two groups on all relevant dimensions. Importantly, participant selection was carried out prior to hypothesis testing.

3.2.2 Experimental paradigm

The experimental procedure was exactly the same as in chapter two.

3.2.3 Statistics

Statistical inference was conducted similarly to analysis of chapter two. To investigate if metacognitive efficiency was different between the autism and comparison group, I fitted a linear model with $Mratio$ from a single-subject fit as dependent variable, clinical group [autism: -0.5, comparison: 0.5] and the covariates (standardized age, IQ, gender [-1: female, 1: male] and education (edu) [1: no education, 2: high school or equivalent, 3: some college, 4: BSc, 5: MSc, 6: doctoral]) as independent variables:

$$\begin{gathered} log(Mratio)_s \sim \beta_0 + \beta_1 group_s + \\ \beta_2 age_s + \beta_3 IQ_s + \beta_4 gender_s + \beta_5 edu_s + \varepsilon_s \end{gathered} \tag{3.1}$$

I also conducted hierarchical regressions using the HMeta-d toolbox in which $Mratio$ in the autism and comparison groups were estimated in separate models that controlled for the following covariates:

$$\begin{gathered}log(Mratio)_s \sim \beta_0 + \beta_1 age_s + \\ \beta_2 IQ_s + \beta_3 gender_s + \beta_4 edu_s + \varepsilon_s \\ where\ s = 1,\ \ autistic\ participants\end{gathered} \tag{3.2}$$

$$\begin{gathered}log(Mratio)_s \sim \beta_0 + \beta_1 age_s + \\ \beta_2 IQ_s + \beta_3 gender_s + \beta_4 edu_s + \varepsilon_s \\ where\ s = 1,\ \ comparison\ participants\end{gathered} \tag{3.3}$$

To assess significance I computed the probability P_θ of overlap between the HDI posterior distribution of Mratio in the autism and comparison group:

$$P_\theta(Mratio_{autism} < Mratio_{comparison}) \tag{3.4}$$

To assess whether the effect of logRT on confidence was different for autistic and comparison participants, I conducted hierarchical mixed-effect regression models using the “lme4” package in R (version 1.2.5033) similar to the method used in chapter two, but now using a dummy variable denoting clinical group (group [autism: -0.5, comparison: 0.5]) instead of continuous autistic trait scores. To visualize the direction of significant effects I obtained the beta-coefficients of logRT on confidence for each clinical group and on error and correct trials, separately:

$$\begin{gathered}conf_{acc/group} \sim \beta_0 + \beta_1 logRT_s + \beta_2 gender_s + \\ \beta_3 edu_s + \beta_4 IQ_s + \beta_5 age_s + \varepsilon\end{gathered} \tag{3.5}$$

Table 3.1: ***Table 3.1. Demographic differences between participant groups.*** *Group mean and standard error from the mean is given for age in years, % female, a proxy of IQ derived from the International Cognitive Ability Resource [Condon and Revelle, 2014] and education [1: no education, 2: high school or equivalent, 3: some college, 4: BSc, 5: MSc, 6: Doctoral]. All comparisons are non-significant for independent samples t-test contrasts between groups (assuming equal variances).*

Demographic	**ASD**	**Comparison**
age	37.90 ± 1.59	35.50 ± 2.10
female (%)	93	85
IQ	9.36 ± 0.54	7.90 ± 0.52
education	4.00 ± 0.06	3.92 ± 0.19

3.3 Results

3.3.1 Performance and validation checks

As a result of the selection procedure described in **Methods**, age (M_{autism} = 37.90 (SE = 1.59), $M_{comparison}$ = 35.50 (SE = 2.10), independent samples t-test, t_{78} = 0.90, P = 0.37), gender (M_{autism} = 93%, $M_{comparison}$ = 85%, independent samples t-test, t_{78} = 1.07, P = 0.29), education (M_{autism} = 4.00 (SE = 0.06), $M_{comparison}$ = 3.92 (SE = 0.19), independent samples t-test, t_{78} = 0.25, P = 0.80) and IQ scores (M_{autism} = 9 (SE = 0.54), $M_{comparison}$ = 7.90 (SE = 0.52), independent samples t-test, t_{76} = 1.45, P = 0.15), were not statistically different between the autism and comparison group (see **Table 3.1**).

Next, I conduct some validation checks on the metacognitive and mentalizing metrics. The main variables of interest were normally distributed and visually similar to the results in chapter two: first order performance (M = 74.34% ± 0.006; W = 0.98, P = 0.12; **Figure 3.1a**) and $Mratio$ (M = 0.653 ± 0.045; W = 0.987, P = 0.60; **Figure 3.1b**). As a result of the calibration procedure, first order performance was not statistically different between the autism (M = 0.75 ± 0.01) and comparison groups (M = 0.74 ± 0.008; equal variances: P = 0.73, K = 0.15; independent samples t-test, t_{78} = 0.519, 95% CI = [-0.019, 0.032], P = 0.61; **Figure 3.1c**). Finally, I averaged the log of response times (logRT) across trials of the metacognition task for each subject and plotted the distribution in **Figure 3.1d**). Average logRT in the autism group (M= -7.39e-17 ± 6.05e-17) and in the comparison group (M= -2.59e-17 ± 6.17e-17) were not statistically different (t_{71} = 0.49, 95% CI = [-2.43, 1.47], *P* = 0.63).

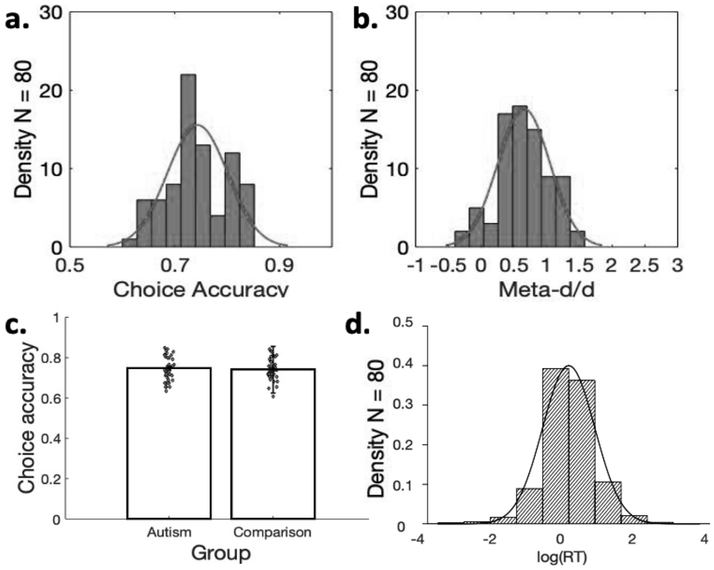


Figure 3.1: ***Choice accuracy on the metacognition task.*** ***a.*** *Histogram distribution of choice accuracy on the metacognition task in the group as a whole (N=80).* ***b.*** *Histogram distribution of metacognitive efficiency (meta-d'/d') on the metacognition task in the group as a whole (N=80).* ***c.*** *Average choice accuracy was matched for autism (N=40) and comparison participants (N=40) on the metacognition task. Error bars represent group mean ± SEM.* ***d.*** *Histogram distribution of the log of standardized response times (logRT) on the metacognition task in the group as a whole (N=80).*

I again sought to ensure key variables related to metacognition and mentalizing were independent of first order perceptual task performance. Staircase variability, the ratio of the standard deviation and the mean dot difference, was not correlated with mentalizing efficiency (rs_{78} = -0.044, *P* = 0.71; **Figure 3.2a**) and was not statistically different between groups (95% CI = [-0.036, 0.026], t_{78} = -0.31, *P* = 0.756; **Figure 3.2b**). In addition, staircase variability was not correlated with metacognitive efficiency (rs_{78} = 0.031, *P* = 0.782; **Figure 3.2c**). Perceptual sensitivity (*d'*) was not correlated with mentalizing efficiency (rs_{78} = 0.011, *P* = 0.924; **Figure 3.2d**) and was not statistically different between groups (95% CI = [-0.083, 0.309],

$t_{78} = 1.15$, $P = 0.253$; **Figure 3.2d**).

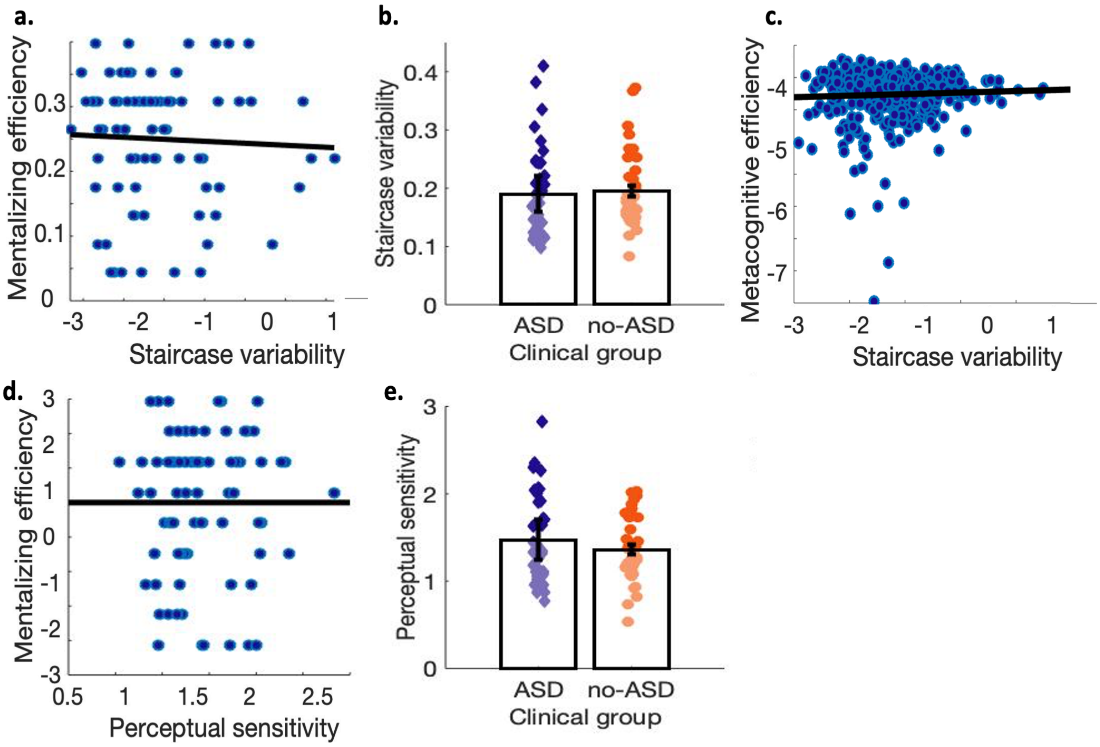


Figure 3.2: ***Correlations between the main variables of interest. a.*** *Mentalizing efficiency and staircase variability in the sample as a whole (N=80) were not correlated.* ***b.*** *Staircase variability was not different between the autism (N=40) and comparison groups (N=40).* ***c.*** *Metacognitive efficiency (meta-d'/d') and staircase variability in the sample as a whole (N=80) were not correlated. d. Mentalizing efficiency and perceptual ability (d') in the sample as a whole were not correlated.* ***e.****Perceptual ability (d') was not statistically different between autism (N=40) and comparison participants (N=40). Error bars represent the group means ± SEM.*

3.3.2 Posterior predictive checks

Finally, I asked whether the two HMeta-d models were reliable by means of convergence checks and posterior predictive checks. The hierarchical regression model converged well, indicated by the Gelman-Rubic statistics ($\hat{R}_{Mratio}$=1.0001 and plotted chains in **Figure 3.3a**). In addition, posterior predictive plots recaptured key patterns of the participants' confidence responses correctly **Figure 3.3b**). The same was true for separate model fits to the comparison group ($\hat{R}_{Mratio}$=1.0014; **Figure 3.3c**) and autism group ($\hat{R}_{Mratio}$=1.002; **Figure 3.3d**).

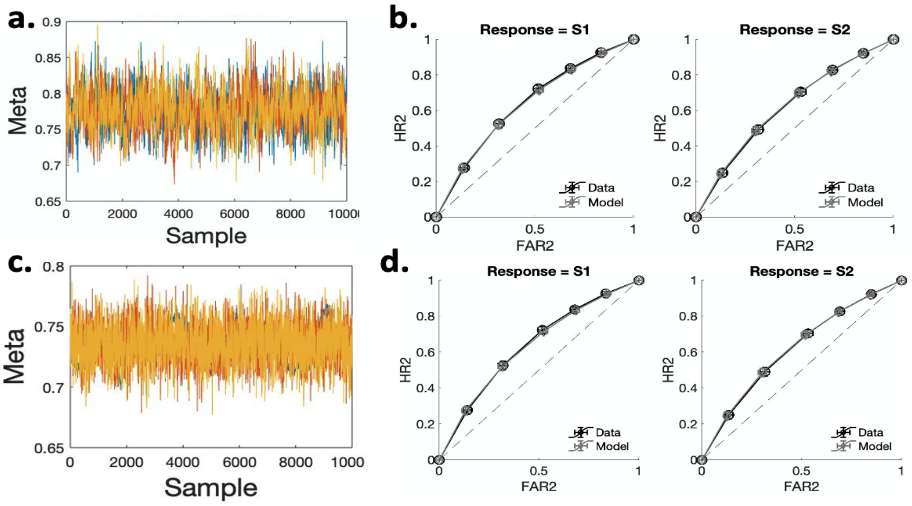


Figure 3.3: ***Posterior predictive checks on HMeta-d fits. a.*** *MCMC chains for parameter meta-d'/d' (metacognitive efficiency) from the hierarchical regression model on autistic participants (N = 40)* ***c.*** *and for comparison participants (N = 40).* ***b.*** *Observed and model estimates for the Type II ROC curves for leftward (S1) and rightward (S2) responses from the hierarchical regression model are plotted for autistic participants (N = 40)* ***d.*** *and for comparison participants (N = 40). Error bars represent the mean ± standard error of the mean.*

3.3.3 Autism as a metacognitive condition

Having shown that the two groups were matched in terms of demographics and general cognitive ability, I next asked if autistic participants had lower mentalizing efficiency than comparison participants by testing a linear regression model with mentalizing efficiency as independent variable and clinical group [autism: -0.5, comparison: 0.5] and the covariates (age, gender, IQ and education) as predictor variables. When I do this, I find that mentalizing efficiency was indeed lower for autistic participants than comparison participants, but not significantly so (linear regression: β_{group} = -0.43 (0.25), t_{68} = -1.72, P = 0.089).

Next, I use a similar linear regression model to test if the autism group had lower metacognitive efficiency than the comparison group. In line with my pre-registered hypotheses, this indeed revealed significantly lower metacognitive efficiency in autistic participants than in comparison participants (linear regression model: β_{group} = -0.60 (0.25), t_{63} = -2.46, P = 0.016; **Figure 3.4a**) with no effects of the covariates. I next estimated metacognitive efficiency within a hierarchical model fitted to each group separately, while accounting for the effects of IQ, age, gender, education. The 95% HDI is lower for the autism group than for the comparison group. When I do this, the HDI of metacognitive efficiency in the autism group (HDI [0.92, 0.55]) was quantitatively lower than that of the comparison group (HDI [0.84, 0.52]) in 78% of the samples $P_{\theta(ASD\ <\ comparison)}$= 0.78; **Figure 3.4b**), although did not reach significance at the classical 95% threshold. Taken together, these analyses provide some evidence in support of my pre-registered hypothesis of lower metacognitive efficiency in autism.

Finally, building upon a hierarchical mixed-effect regression model of trial-by-trial predictions of confidence on the metacognition task, I next tested whether the model could better predict confidence levels when the predictors (**Equation 2.4**), were allowed to vary as a function of whether the subject was autistic or not (**Equation 2.5**). A likelihood ratio test indicated that this was the case ($\chi^2(4)$ = 966.46, $P <$ 2.20e-16) which was further strongly confirmed by goodness-of-fit indices (ΔLL = -484, ΔAIC = 958, ΔBIC = 929 and $\Delta Deviance$ = 966), supporting the prediction that confidence formation in autistic participants is qualitatively distinct to comparison participants.

Consistent with the results of Experiment 1 I found that autistic participants report lower confidence than comparison participants in general (hierarchical regression model, main effect of group: $\chi^2(1)$=768.50, $P <$ 2.0e-16, β=0.82, SE = 0.03). Autistic participants show a marginally lower impact of response times in error trials than comparison participants (three-way interaction logRT x group x accuracy: $\chi^2(1)$ = 3.086,P = 0.060, β = 0.10, SE = 0.06). In **Figure 3.4c** I plot the impact of response times on confidence on error and correct trials separately, which shows that the negative impact of RT on confidence was less negative in autistic participants than in comparison participants, suggesting a weaker influence on response times on confidence in error trials.

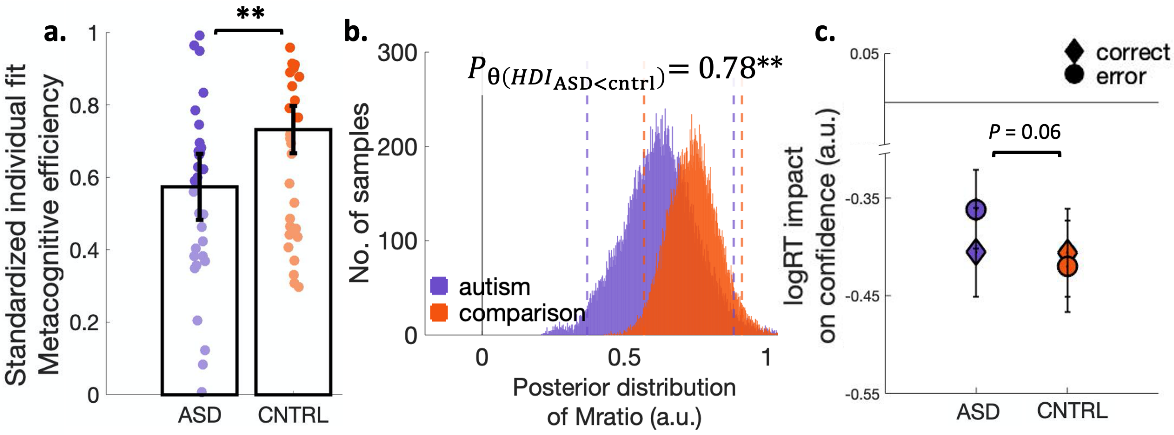


Figure 3.4: ***Differences in metacognitive efficiency and confidence formation in autism.*** ***a.*** *Metacognitive efficiency estimated from a single-subject Bayesian model fit is significantly lower in the autism group (N=40) than in the comparison group (N=40). Error bars represent group mean ± SEM.* ***b.*** *Posterior estimates of metacognitive efficiency from independent group model fits (autism in purple, controls in orange) where the dashed lines represent the highest density intervals (HDI) and* P_θ *represents the probability that the HDI of the autism group is lower than the HDI of the comparison group.* ***c.*** *Impact of logRT on confidence on error and correct trials for autism and comparison participants. Error bars represent group means ± SEM.*

3.4 Discussion

In this chapter, I leveraged a model-based approach to examine the interplay between metacognitive efficiency and mentalizing efficiency in autism. Using hierarchical regression models I found that metacognitive efficiency is compromised in autism, and reveal a weaker association between response times and confidence in autistic deciders in contrast to matched controls. These results are in line with some earlier findings that metacognition is compromised in autism [Grainger et al., 2016a, Nicholson et al., 2020, Williams et al., 2018] but not with other studies [Carpenter et al., 2019b, Wojcik et al., 2013]. This inconsistency may be driven by metacognitive metrics being confounded by differences in first-order ability, such as choice accuracy and confidence [Fleming and Lau, 2014, Masson and Rotello, 2009]. This confound is particularly pertinent to the question at hand, as perceptual sensitivity and confidence may vary across autistic and comparison populations [Milne et al., 2002, McMahon et al., 2016, Zalla et al., 2015].

It could be, for example, that previous estimates of metacognitive efficiency in the autism group were higher because of a general tendency for under-confidence in the autism group. By using a signal detection framework to calculate metacognition [Maniscalco and Lau, 2012, Maniscalco and Lau, 2014] I ensured that such confounds could not have affected these current results. The hierarchical inference technique used here to estimate metacognitive efficiency further has as benefit that it takes into account both within- and between subject variability. This is relevant, as autistic people have been reported to have greater variation in their cognitive abilities [Joseph et al., 2002, Doyle, 2020]. Indeed, I found that autistic participants had marginally higher variance in their metacognitive profile than comparison participants (KS = 0.275, P = 0.079; **Figure 3.4a**).

Even though this study went above and beyond to ensure a metric of metacognitive efficiency was isolated from differences in first-order performance, the current approach still leaves open the possibility that the used metric of mentalizing efficiency was not tested in a fully controlled manner. It could be, for example, that the ability to learn stimulus response associations have confounded my used metric of mentalizing efficiency [Livingston et al., 2019b]. I here tried to circumvent this potential artefact by using participants' responses to questions about the feelings of the figures, which cannot easily be learnt, and by relying on a multiple-choice design rather than an open response, which is less confounded by differences in verbal fluency [White et al., 2011]. Future studies should go beyond these steps and develop computational and empirical methods to separate mentalizing efficiency from first-order cognition, as has been done in the measurement of metacognitive efficiency.

Another limitation of this study is that of domain-generality. There is reason to believe that metacognitive efficiency measured from perceptual decision-making is similar to metacognitive efficiency measured in other domains, such as from mnemonic or numerical decision-making tasks [Bronfman et al., 2015, Talluri et al., 2018, van der Plas et al., 2022]. However, other studies found selective differences in perceptual metacognition between groups, in the absence of differences in memory metacognition [Fleming and Lau, 2014]. The possibility of dissociations between domains suggests an unlikely, albeit possible, chance that mentalizing efficiency is only related to metacognitive efficiency when the latter is measured in the context of a perceptual task. Future studies should test the interplay between metacognition and mentalizing across a wider range of cognitive domains.

The current results indicate that autism may be considered a condition that is characterized by difficulties with both metacognition and mentalizing, which is in line with the classical notion of autism constituting a general form of 'mind blindness' [Happé, 1994, Happé, 2003, Baron-Cohen et al., 1985]. Even if this notion has mostly been supported with findings of mentalizing difficulties in autism [Livingston et al., 2019b, White et al., 2011, White et al., 2009, Baron-Cohen, 1992, Abell et al., 2000], more recent work has also started to investigate metacognitive efficiency in autism [Nicholson et al., 2020, Nicholson et al., 2019, Carpenter et al., 2019b]. This avenue of new research may be particularly important, given that new work is highlighting the relevance of metacognitive processes on a variety of other cognitive processes.

In highly controlled, experimental paradigms where the underlying experimental structure is consistent and predictive conditions being primarily metacognitive in nature may often go unnoticed, as has been reported in the case of autism [Livingston et al., 2019b]. A metacognitive explanation of autism may explain why everyday decision-making scenarios seem more compromised in autism that would be expected on the basis of autistic people's performance on structured tasks [Geurts et al., 2020, Luke et al., 2012], and why decision-making difficulties in autism are more pronounced at, or limited to, the explicit (verbal) level, such as the way in which choice preferences are reflected upon or evaluated, but much less to the (implicit) process of deciding [Chantiluke et al., 2015, D'Cruz et al., 2016, Johnson et al., 2006, Larson et al., 2011, McPartland et al., 2012, Solomon et al., 2015]. Everyday instances of (social) decision-making are ambiguous and convoluted, and often provide unreliable feedback or information [Jänsch and Hare, 2014, Robic et al., 2015] and, thus, rely to a greater extent on metacognition than lab-based tasks. These results suggest that it is exactly this feature of everyday decision-making that autistic people struggle most with. These results are in line with those of a recent literature review [Van der Plas, Mason et al., 2022], and predict that metacognitive difficulties in autism may explain some of the real everyday struggles autistic people face with decision-making [Koren et al., 2006, Livingston et al., 2019b].

In summary, in this chapter I show that metacognitive efficiency in compromised in autism. Specifically, I show that the metacognitive difficulties in autism may be driven by a lower association between response times and confidence in autistic deciders in contrast to age, gender, IQ and education matched comparisons. In the General Discussion I will review these results in light of the mentalizing-is-prior hypothesis.

Chapter 4

Cultural contributions to metacognition

4.1 Introduction

In the general introduction I argued that aspects of explicit metacognition may be culturally acquired and determined by the extent to which cultures place emphasis on discussing and understanding the mental states of self and other [Cleeremans et al., 2020, Heyes et al., 2020, Heyes and Frith, 2014]. In other words, just as children learn to understand the meaning of written words from teachers and parents, children who grow up in cultures where working together is the norm may develop a stronger awareness of their own and others' mental states.

A key implication of this cultural origins hypothesis is that metacognition should be subject to cultural variation to the extent that there are cultural differences in social collaboration and integration. Specifically, the supra-personal functions of metacognition–accurate communication and broadcast of private mental states to others–should have benefits not only to the owner of those skills, but also to other members of the social group with whom they make decisions and coordinate action. Consequently, it is in the interests of a person with enhanced metacognitive skills to teach those skills, deliberately or inadvertently, to others in the group. The requirement to do so is presumably stronger in more socially integrated groups, such as cultures where collaboration and shared goals are more common. The cultural origins hypothesis suggests that these slight differences in the importance of communication may have a downstream impact on objectively measured metacognitive abilities (the alignment between confidence and performance).

A rich source of potential cross-cultural differences in social integration has been documented in studies comparing China with the West. Chinese populations are more likely to pay attention to and conform to others' opinions than UK or US populations [Korn et al., 2014, Mesoudi et al., 2015, Oeberst and Wu, 2015]; are thought to be more interdependent than independent in thinking styles [Singelis, 1994]; and be more collectivist in emphasizing harmony with others than Western countries [Hofstede, 2011, Markus and Kitayama, 2010, Weber, 1905]. However, whether cultural background similarly affects explicit metacognition remains unknown. Here, by applying recently developed psychophysical tools for isolating and quantifying the capacity for explicit metacognition about simple decisions, I seek to evaluate this hypothesis.

Previous cross-cultural studies of metacognition have focused on quantifying differences in subjective estimates of confidence. For example, a typical study might ask subjects general knowledge questions such as *"Which one is further north: New York or London?"* after which participants indicate their confidence that the decision was correct. Such studies have found that Chinese populations report higher confidence than US or UK populations [Moore et al., 2018, Yates et al., 1998, Yates et al., 1989]. It is important to note, however, that average confidence is only one facet of metacognition, known as metacognitive bias, and can vary independently of metacognitive sensitivity, the ability to discriminate between correct and error trials using confidence ratings [Fleming and Lau, 2014, Maniscalco and Lau, 2012, Maniscalco and Lau, 2014]. In other words, a highly confident person may still realize when they are wrong, and rate lower confidence accordingly—thus demonstrating good metacognitive sensitivity. This capacity for metacognitive sensitivity, rather than idiosyncrasies in metacognitive bias, is also likely to be the key variable for effective collaboration with others [Bahrami et al., 2010, Fusaroli et al., 2012].

Two previous studies have quantified cross-cultural differences in both metacognitive bias and sensitivity. Yates and colleagues found that, despite a heightened (overconfident) metacognitive bias, metacognitive sensitivity was also higher in Chinese than US populations, as measured by probability judgment discrimination scores [Yates et al., 1989]. Another study also found heightened metacognitive bias in Chinese people living in Taiwan in comparison to Japanese and American populations, but inconsistent effects on metacognitive sensitivity [Yates et al., 1998]. However, in both of these studies, first-order performance (judgment accuracy) was left free to vary across a wide range, and differences in metacognitive sensitivity could be confounded by group differences in accuracy [Fleming and Lau, 2014]. This is an important limitation, given that metacognitive sensitivity is affected by choice accuracy—people tend to discriminate better between their errors and correct decisions when the task at hand is easier. Moreover, both of these studies looked at associations between average confidence and average accuracy collapsed over groups of trials. Much less is known about cultural differences in the computational processes that give rise to subjective confidence judgments. For instance, recent work in this area characterized how evidence accumulation may continue for a short time after the choice, supporting endogenous error monitoring [Murphy et al., 2015, Rabbitt, 1966] and changes of mind [Resulaj et al., 2009, van den Berg et al., 2016]. More recently this line of work has been extended to ask how exogenous evidence presented after an initial choice may lead to later changes of mind. In these tasks, participants first make a judgment based on some evidence (e.g., an estimation of the direction of a random-dot motion display), after which they are presented with new evidence (e.g., additional motion) and are asked to make a final judgment. The general finding is that people tend to update their final judgment after seeing the new evidence and that this updating is stronger when the new evidence is more reliable or stronger

[Bronfman et al., 2015, Fleming et al., 2018], or when the new evidence confirms the initial judgment (known as 'confirmation bias'; [Talluri et al., 2018]). Given the central role that post-decision processing plays in promoting openness to others' (conflicting) viewpoints [Schulz et al., 2020, Rollwage et al., 2018], it could be that cultural norms of harmony and collaboration selectively impact metacognition through shaping the processing of post-decision evidence. This hypothesis selectively predicts modulations in the computational processes that update confidence estimates based on new evidence. One interesting prediction of the cultural origins hypothesis of metacognition is that any cultural difference should be relatively domain-general, because the skills that are being acquired are metacognitive in nature rather than how to handle a particular type of information [Heyes et al., 2020].

By leveraging frameworks derived from psychophysics and signal detection theory, it has now become possible to isolate precisely metrics of metacognitive efficiency in laboratory tasks from the extent to which subjects recognize their mistakes by adjusting their sensitivity to new evidence [Maniscalco and Lau, 2012, Galvin et al., 2003, Maniscalco and Lau, 2012]. Building upon these recent methodological advances [Fleming and Dolan, 2012, Fleming and Lau, 2014, Frith and Frith, 2012, Yeung and Summerfield, 2012] I sought to provide an initial assessment of whether metacognitive processes engaged in processing post-decision evidence differ between individuals drawn from distinct Northern European and Chinese cultural milieus. In a post-decision evidence task, new information about the stimulus is given after an initial decision, but before the confidence estimate [Fleming et al., 2018]. Intuitively, this new evidence allows individuals to update their belief in the accuracy of their initial choice—providing an empirical window into the process of confidence formation. Given that confidence estimates are thought to generally involve some form of post-decision evidence (reflective) processing [Navajas et al., 2016], exogenously manipulating post-decision evidence provides good experimental control over this process [Rollwage et al., 2018].

After an initial perceptual decision about the direction of a patch of randomly moving dots (left versus right), participants were shown additional (post-decision) evidence and asked to rate their confidence that the initial decision was correct. Using a calibration procedure, I selected stimuli of similar perceptual strength across individuals and sites to match first-order task difficulty, such that any difference in metacognition between cultures was unrelated to the first-order performance. To ensure well matched samples, I compare the profiles of confidence judgments in Chinese and British samples matched for occupation (full-time students at Peking University, PKU, and University College London in the UK), age, gender, income and IQ. I only recruited Chinese/British citizens that had at least one parent that was born and raised in mainland China/Britain and had not lived more than one year abroad.

The results show that the sample with a Chinese cultural background showed distinct metacognitive profiles when compared to those from the British cultural milieu. Specifically, they showed a heightened sensitivity to post-decisional evidence, leading them to (correctly) change their minds more about errors compared to those from the UK. Taken together, these results reveal heightened metacognitive sensitivity and post-decisional processing in samples from the Chinese cultural background in the absence of differences in first-order perceptual performance, consistent with the hypothesis that cultural variation contributes to metacognition.

4.2 Methods

4.2.1 Participants

I recruited N = 83 participants at both Peking University (PKU) in Beijing, China and University College London (UCL) in London, UK (**Table 4.1**). I did not conduct statistical tests to predetermine the sample size, as the effect size for a potential cultural difference was unknown. Instead, I used a sample size similar to those used in previous publications using the same paradigm [Fleming et al., 2018]. At both sites the experiment was advertised via an online platform and flyers on campus, from which I recruited participants that were: (1) full-time students at PKU/UCL; (2) Chinese/British citizens; (3) had at least one parent that was born and raised in mainland China/Britain; and (4) had not lived more than one year abroad. All participants had normal or corrected-to-normal vision and no history of neurological or psychiatric illness. Instructions, advertisements and questionnaires in English were translated to Mandarin Chinese and then back translated by an independent translator. The study was approved by the University College London Ethics Committee (1260/003) and by the Ethics Committee of School of Psychological and Cognitive Science at Peking University. All participants gave written informed consent before taking part in the experiment.

Exclusion criteria were defined a priori and are the same as the exclusion criteria employed by several previous papers using the same or similar tasks [Rollwage et al., 2018, Fleming et al., 2018]. Two participants were excluded from the PKU dataset: one participant did not follow task instructions and one participant performed below an a priori accuracy cut-off threshold (i.e., less than 60% accuracy). Three participants were excluded from the UCL dataset: one participant was found not to have met the recruitment criteria after data collection (not a full-time student), one participant lacked variability in their confidence ratings (881/900 trials were rated as 100% confident) and one participant performed below the accuracy cut-off threshold of 60%. This resulted in the analysis of thirty-nine participants per site (N = 78 participants in total of which 39 females, mean age: 22.63 ± 0.33 years). All key site differences reported in the Results section remained significant after I re-introduced these participants.

In addition, I re-analyzed an original dataset [Fleming et al., 2018] which was collected as the first part of a two-day study at New York University (NYU). This dataset consisted of N = 25 participants (14 female, mean age: 24.0 ± 0.72 years) but information on the cultural background of the sample was not collected, thus I cannot be sure that these participants formed an adequate comparison group. The NYU recruitment was approved by NYU's University Committee on Activities Involving Human Subjects and all participants provided written consent before taking part in the experiment.

4.2.2 Experimental paradigm

The experiment was programmed in Matlab 2014b (MathWorks) using Psychtoolbox (version 3.0.12) and presented on a desktop monitor at approximately 45 centimetres viewing distance. Stimuli were random dot kinematograms (RDKs): 30 moving dots (0.12 diameter) that appeared in a 7 diameter circular white aperture for 300 milliseconds. The movement of the dots was generated by replotting the dots every three video frames, with a subset moving horizontally to either the left or the right and the remainder moving in a random direction. The subset that moved in the coherent direction was manipulated across conditions as giving rise to weak, medium or strong evidence strength. To ensure that these conditions were perceptually equivalent across participants, I performed a calibration procedure in which I estimated each participants' psychometric function for a broad range of evidence strength levels and then selected the three evidence strength levels that were associated with three pre-specified levels of accuracy (weak = 60%, medium = 75% and strong = 90%).

On the psychophysical task, participants were shown 900 samples of evidence (RDK stimuli, pre-decision evidence) with variable evidence strength and were asked to judge the direction of dot movement (left or right). Participants indicated their choice by pressing a keyboard button [left: 1; right: 2] within 1,500 ms. After the choice, participants were shown "bonus" post-decision evidence where the dots moved in the same direction but with variable evidence strength (weak, medium, strong). In total, there were thus nine experimental conditions in a 3 (three pre-decision evidence strength levels) x 3 (three post-evidence strength levels) factorial design; **Figure 4.2a**). At the end of every trial, participants were asked to rate their confidence that the initial judgment was correct on a scale ranging from 0 to 100%. Participants indicated their response by selecting a point on the scale with the mouse cursor within 3,000 ms. I implemented a Quadratic Scoring Rule (QSR) to motivate participants to report their confidence as accurately as possible. In particular, participants earned maximum points on a trial if they rated the lowest possible confidence about an incorrect judgment, or if they rated the highest possible confidence about a correct judgment. The total duration of the task was approximately 40-50 minutes and participants were instructed to take six self-paced breaks at specific moments in the task.

4.2.3 Additional measures

After the psychophysical task, I administered three additional surveys: Self-Construal scale [Singelis, 1994], Analysis-Holism scale [Choi et al., 2003], and Culture-Free Intelligence test [Cattell, 1943].

The Analysis-Holism scale (AHS) measures individuals' analytical versus holistic thinking tendency [Choi et al., 2007]. People that think more analytically focus more on objects instead of on the whole, and usually desire one true answer instead of accepting that multiple dissimilar or even opposing truths can be valid at the same time. A total of 24 items (4 sub-scales) are rated from 1 ('strongly disagree') to 7 ('strongly agree'). Following a standard coding procedure, I reverse-coded seven items and summed the resulting scores.

The Self-Construal Scale (SCS) measures the strength of individuals' inter-dependent and in-dependent self-construal [Singelis, 1994], i.e., how important people think that maintaining harmony within their social group is. A total of 24 items (corresponding to either the independent or inter-dependent subscale) were rated from 1 ('strongly disagree') to 7 ('strongly agree'). I used an available translation from [Huang et al., 2007] which was back-translated by an in-dependent Chinese-English speaker.

The Cattell Culture-Free Intelligence Test (CFIT) is a non-verbal measure of individual's fluid intelligence that minimizes the influence of verbal fluency, culture and education [Cattell, 1943]. Participants were asked to complete visual puzzles by selecting one of four multiple choice options, which each pertained to four subtasks and had different instructions. I explained the instructions and signaled a pre-defined time limit for each sub-task. Raw accuracy scores were converted to IQ scores following a standard coding table.

A collaborator translated the Analysis-Holism scale and the Culture-Free Intelligence Task to Mandarin Chinese and I used a published translation of the Self-Construal scale [Singelis, 1994]. All Mandarin Chinese translations of the questionnaires were back translated by an independent translator to ensure translation quality before the questionnaires were used at PKU. In **Table4.1**) I report the details of these questionnaires and compare their scores across sites.

4.2.4 Statistics

Group differences were tested with two-tailed independent samples t-tests (assuming equal variances). To assess the effects of my factorial design on accuracy and confidence, I conducted hierarchical mixed-effect regression models using the 'lme4' package in R (version 3.3.3) and plotted the behavioral data and the output of the model fits in MATLAB (version R2018a). I obtained the *P*-values of the regression coefficients using the *car* package. Given that we expected individual differences in the association between confidence and task variables between individuals even within each cultural group, I specified a random effect at the subject level corresponding to each fixed effect of interest. I report type III Wald chi-square tests (χ^2), degrees of freedom (df) for fixed effects, and estimated beta-coefficients (β) together with their standard errors of the mean (± SEM) and *P*-values of the associated contrasts.

I investigated the effect of the pre-decision evidence strength (pre) [weak: -0.5, medium: 0, strong: 0.5] across sites [1: PKU, 2: UCL] on trial-by-trial accuracy [0: error, 1: correct] with the following hierarchical mixed-effect logistic regression model:

$$acc \sim site * pre + (1 + pre|subj) \tag{4.1}$$

To predict confidence, I used a hierarchical mixed-effect regression model with trial-by-trial confidence as the dependent variable, and accuracy [-1: error, 1: correct], z-score of the log response time (RT), pre-decision evidence strength (pre) [weak: -0.5, medium: 0, strong: 0.5], post-evidence strength (post) [weak: -0.5, medium: 0, strong: 0.5], site [1 = PKU, 2 = UCL] and their interactions as predictors:

$$\begin{gathered} conf \sim site * (acc + pre + post + pre * post + pre * acc + post * \\ acc + pre * post * acc + log_{RT}) + (1 + acc + pre + post + pre \\ * post + pre * acc + post * acc + pre * post * acc + RT|subj) \end{gathered} \tag{4.2}$$

After demonstrating that I replicate some previously reported findings [Fleming et al., 2018] in each site separately, I combined the two datasets and included a site interaction term to investigate whether the effects are consistent between PKU and UCL (see **Figure 4.3**) for a comparison of all three sites including NYU). To investigate whether the model's prediction of confidence improved when cross-cultural terms were included, I conducted a Likelihood Ratio Test that assesses the benefit of including interactions with site, here expressed in terms of the Akaike Information Criterion (AIC): $\Delta AIC = AIC_{nosite} - AIC_{withsite}$, and the Log Likelihood (LL): $\Delta LL = LL_{withsite}$–$LL_{nosite}$ with associated P value extracted from a type III Wald chi-square tests ($\chi 2$). In addition, I confirmed that simulating data from the summary statistics of the hierarchical regression model in **Equation 4.2** successfully recaptured key features of the actual dataset (**Figure 4.4b**).

To visualize the direction of the effects in **Equation 4.2**, I obtained the beta-coefficients of the pre-decision evidence conditions (pre) [weak: -0.5, medium: 0, strong: 0.5] and the post-evidence conditions (post) [weak: -0.5, medium: 0, strong: 0.5] and their interactions on confidence for each site [1: PKU, 2: UCL] and on error and correct trials separately:

$$\begin{aligned} confidence_{err/corr} \sim pre + post + pre * post + RT \\ + (1 + pre + post + \ pre * post + RT \mid subj) \end{aligned} \tag{4.3}$$

4.3 Results

I analysed the data of N = 78 participants (N = 39 at each site) who were matched in terms of age (M_{PKU} = 22.33 (SE = 0.38), M_{UCL} = 22.92 (SE = 0.54), independent samples t-test, t_{76} = -0.89, 95% Confidence Interval (CI) = [-1.91, 0.73], P = 0.38), gender (M_{PKU} = 49%, UCL = 51%, t_{76} = -0.22, 95% CI = [-0.25, 0.20], P = 0.82) and annual family income (their parents combined gross income before tax, converted from Chinese renminbi (¥) to pounds (£) at 2017 purchasing power parity) relative to the *per capita* purchasing power parity at the time of recruitment (M_{PKU} = £37,615.38 (SE = 4,535.01) and UCL (M_{UCL} = £39,381.35 (SE = 3,962.23), t_{75} = -0.29, 95% CI = [-13852, 10320], P = 0.77). In addition, I administered a non-verbal measure of fluid intelligence which minimizes the influence of verbal fluency, culture and education (Cattell Culture-Free Intelligence test; [Cattell, 1943]), which showed no differences in general intelligence between both sites (M_{PKU} = 102.36 (SE = 1.79), and M_{UCL} = 101.15 (SE = 1.52), t_{73} = 0.51, 95% CI = [-3.55, 5.96], P = 0.60; see **Table 4.1**) for additional measures).

Table 4.1: ***Demographical and trait differences between PKU and UCL datasets.*** *Group mean ± standard error from the mean. Income in pounds (£) is given relative to the purchase power parity (PPP) ratio between UK and China at the time of recruitment (ratio 1:1.71). Mean composite score ± standard error from the mean is given for the CFIT (Cattell Culture-Free Intelligence Test), AHS (Analysis-Holism scale), ind (independent) and int (interdependent) SCS (Self-Construal Scale). *P $<$ 0.05, ** P $<$ 0.01, *** P $<$ 0.001 for the independent samples t-test contrast between sites (assuming equal variance), bold values indicate a significant group difference.*

Trait	**PKU**	**UCL**
age	22.33 ± 0.38	22.92 ± 0.54
% female	49%	51%
income(£)/PPP	37,615 ± 4,535	39,381 ± 3,962
CFIQ	102.36 ± 1.79	101.15 ± 1.52
AHS	**4.79 ± 0.07**	**5.02 ± 0.07 ***
SCS-ind	**4.72 ± 0.13**	**5.12 ± 0.13 ***
SCS-int	4.61 ± 0.13	4.73 ± 0.14

Before the main task, participants were shown 240 random dot motion stimuli and had to judge the direction of the movement (left, right) without making a confidence estimation. The coherence of dot movement was manipulated across six coherence levels: 3%, 8%, 12%, 24%, 48% and 100%. Participants heard auditory feedback that signaled the accuracy of their judgment (high-pitched tone signaled a correct judgment and low-pitched tone signaled an error judgment). For every participant, a cumulative normal psychometric function was fitted to the data and the three coherence levels that resulted in 60%, 75% and 90% accuracy were used in the main task. In **Figure 4.1a**) I plot the likelihood of participants' rightward judgement across each level of rightward motion coherence (ranging from 100% coherence left to 100% coherence right). Performance during the calibration phase was 76.6% correct (SE = 0.01) in the PKU sample, 75.7% correct (SE = 0.01) in the UCL sample and 73.5% correct (SE = 0.01) in the NYU sample. Using independent samples t-tests, I show that performance on the calibration phase was not different between PKU and UCL (t_{76} = 0.86, P = 0.39) or between UCL and NYU (t_{62} = 1.54, P = 0.13), but that it was higher in PKU than NYU (t_{62} = 2.40, P = 0.02, uncorrected). These performance levels were successfully reflected in the evidence strength levels that the participants received on the main task. The coherence levels of the weak, medium and strong evidence levels were [0.08, 0.21, 0.40] for PKU (average: M = 0.22 (SE = 0.01), [0.10, 0.26, 0.46] for UCL (average: M = 0.27 (SE = 0.02) and [0.13, 0.34, 0.56] for NYU (average: M = 0.34 (SE = 0.03). As a result of this, first-order performance in the main experiment was matched across sites (**Figure 4.1b**).

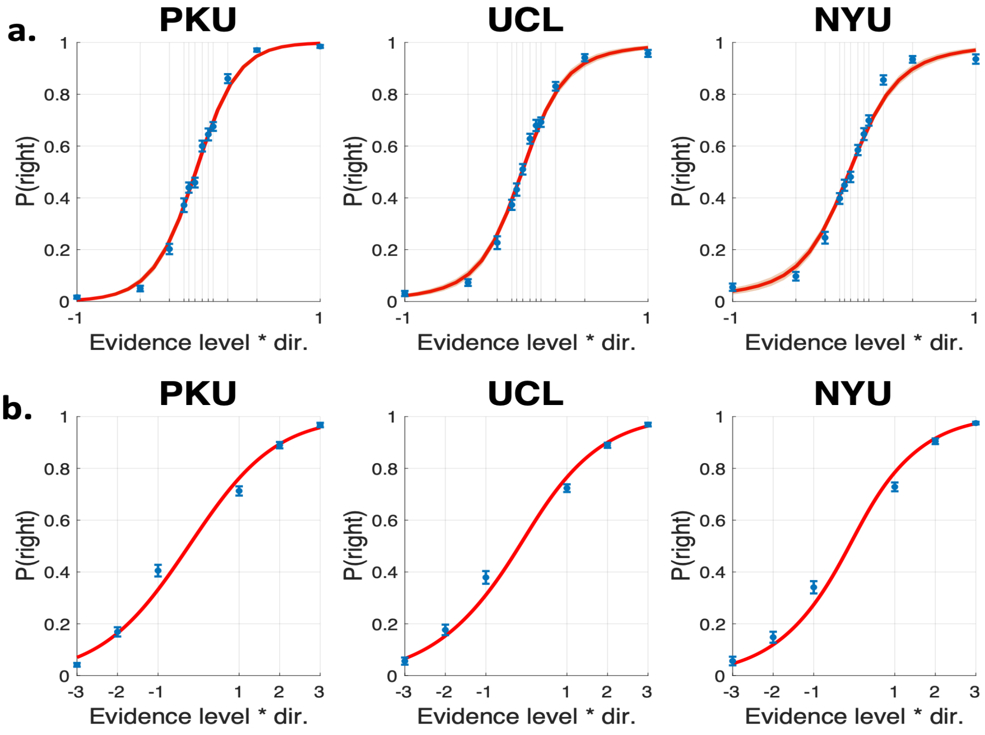


Figure 4.1: ***First-order performance across sites. a.*** *Probability of choosing right 'P(right)' on the calibration task as a function of six coherence levels multiplied by their direction (dir: 100% left = -1, 100% right = 1).* ***b.*** *Fitted cumulative normal psychometric function (red) and behavioral data (blue) of the probability of choosing rightward direction on the main task as a function of the three coherence levels (1 = weak, 2 = medium, 3 = strong) multiplied by their direction (dir: left = -1, right: 1). Solid lines represent the predictions from the model, dots represent the group mean ± standard error.*

I next turn to the psychophysical task of Experiment 1 (**Figure 4.2a**). As a result of the calibration procedure, the accuracy of participants' initial decision (first-order performance) was not statistically different between sites (M_{PKU} = 83% (SE = 0.01), M_{UCL} = 83% (SE = 0.01), independent samples t-test, t_{76} = -0.20, 95% CI

= [-0.03, 0.02], $P = 0.85$).

Using a hierarchical logistic regression to predict trial-by-trial accuracy, I found that first-order performance was indeed more accurate with stronger evidence (hierarchical linear regression, main effect of pre-decision evidence: $\chi^2(1) = 363.02$, $P < 2e\text{-}16$, β= 2.92 (SE = 0.15), = 19.05, $P < 2e\text{-}16$). As expected, this effect did not interact with site (interaction between site and pre-decision evidence: $\chi^2(1) = 0.94$, $P = 0.33$, β= -0.21 (SE = 0.21), = -0.97, $P = 0.33$; **Figure 4.2b**).

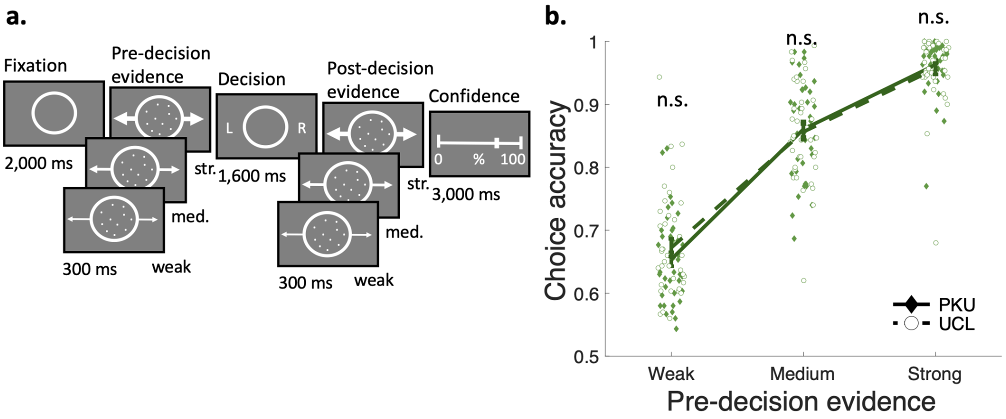


Figure 4.2: ***Task design and matched first-order performance. a.*** *Participants made judgments about the direction (left versus right) of random dot motion. After seeing this pre-decision evidence, participants were shown additional post-decision evidence in the same direction as the pre-decision evidence but of potentially differing strength. Finally, they were asked to rate their confidence of their initial decision being correct on a scale from 0% to 100%, with percentages indicating probability of being correct.* ***b.*** *Choice accuracy was matched between sites (n.s.) and higher following stronger pre-decision evidence levels (P < 0.001, N = 39 participants at each site). Error bars represent group mean ± SEM.*

Having shown that I matched choice accuracy (first-order performance) across sites, the next question was whether confidence ratings varied as a function of the strength of confirming or disconfirming post-decision evidence (weak, medium or strong) that each participant received. Participants were instructed that the new evidence moved towards the same direction as the initial evidence and that they could use both pieces of evidence to rate their confidence about their initial response on a scale from 0 to 100%. I crossed three levels of pre-decision evidence strength with three levels of post-decision evidence strength to create a fully factorial 3 (pre-decision evidence strength) x 3 (post-decision evidence strength) factorial design **Figure 4.2a**).

Across both sites, I replicated key patterns of confidence modulation reported previously [Fleming et al., 2018]: stronger post-decision evidence after an incorrect choice led to lower confidence (as participants could use the new evidence to realise that they were wrong), whereas stronger post-decision evidence after a correct choice led to higher confidence (as participants could use the new evidence to confirm that they were correct):

PKU participants reported higher confidence on correct compared to error trials (main effect of accuracy: $\chi^2(1)$ = 261.77, P < 2.2e-16, β = 0.25 (SE = 0.02), and reported higher confidence after seeing stronger pre-decision evidence (main effect of pre-decision evidence: $\chi^2(1)$ = 5.30, P = 0.02, β = 0.03 (SE = 0.01). The direction in which post-decision evidence influenced confidence was dependent on the accuracy of the initial choice (interaction of post-decision evidence x accuracy: $\chi^2(1)$ = 298.99, P < 2.2e-16, β = 0.25 (SE = 0.01). Specifically, receiving stronger disconfirming post-decision evidence on error trials decreased confidence (as participants could use the new evidence to realize that they were wrong), whilst receiving stronger confirming post-decision evidence on correct trials increased confidence (as participants could use the new evidence to confirm that they were correct). This V-shaped pattern is illustrated in (**Figure 4.3a,b**) (blue lines indicate confidence about correct choices and red lines indicate confidence about incorrect choices). Post-decision evidence decreased confidence on error trials more than it increased confidence on correct trials, as indicated by a negative main effect of post-decision evidence on confidence (main effect of post-evidence: $\chi^2(1)$ = 112.02, P < 2.2e-16, β = -0.15 (SE = 0.01).

UCL participants also reported higher confidence on correct compared to error trials (main effect of accuracy: $\chi^2(1)$ = 230.39, P < 2.2e-16, β = 0.23 (SE = 0.02).

Again, there was an interaction between post-decision evidence and the accuracy of the initial judgement (interaction of post-decision evidence x accuracy: $\chi^2(1)$ = 237.11, $P < 2.2e\text{-}16$, β = 0.21 (SE = 0.01). In addition, a negative main effect of post-decision evidence on confidence shows that post-decision evidence decreased confidence more on error trials than it increased confidence on correct trials (main effect of post-decision evidence: $\chi^2(1)$ = 40.29, $P < 2.2e\text{-}10$, β = -0.09 (SE = 0.01).

I next tested whether a hierarchical regression model better predicted trial-by-trial confidence when the predictor variables (pre- and post-decision evidence levels, accuracy, standardized log response time (RT) and their interactions) were allowed to vary across sites. A Likelihood Ratio Test indicated that this was indeed the case (log likelihood (LL): ΔLL = 11 and Akaike Information criteria (AIC): ΔAIC = 5, χ^2(9)= 23.38, P = 0.005). This effect was not driven by a difference in main confidence level across sites (hierarchical regression model, main effect of site: χ^2(1)= 3.55, P = 0.06, β= -3.88 (SE = 0.02), suggesting a significant role of cultural differences in constructing confidence instead.

I next asked how culture modulated the impact of new evidence on confidence by testing which predictor variables interacted with site (UCL, PKU). I found that post-decision evidence had a higher impact on confidence in the PKU dataset than in the UCL dataset (hierarchical linear regression, interaction of post-decision evidence x site: $\chi^2(1)$ = 6.89, P = 0.009, β= 0.05 (SE = 0.02). This effect was most evident on error trials, as shown by the steeper slope in the PKU dataset (**Figure 4.3b**). Indeed, when when I fitted a hierarchical regression model on error trials only, the impact of post-decision evidence on confidence was significantly higher in the PKU dataset than in the UCL dataset (interaction between site x post-decision evidence on error trials:$\chi^2(1)$ = 4.85, P = 0.03, β= 0.08 (SE = 0.04) but not on correct trials:$\chi^2(1)$= 2.40, P = 0.12, β= 0.02 (SE = 0.02); **Figure 4.3b**). However, the three-way interaction between post-decision evidence, accuracy and site did not reach statistical significance when tested within a single hierarchical regression model ($\chi^2(1)$= 2.23, P = 0.14, β= -0.03 (SE = 0.02), $t_{74.04}$ = -1.49, P = 0.13), suggesting an enhanced susceptibility to new evidence in the PKU sample was not necessarily restricted to error trials.

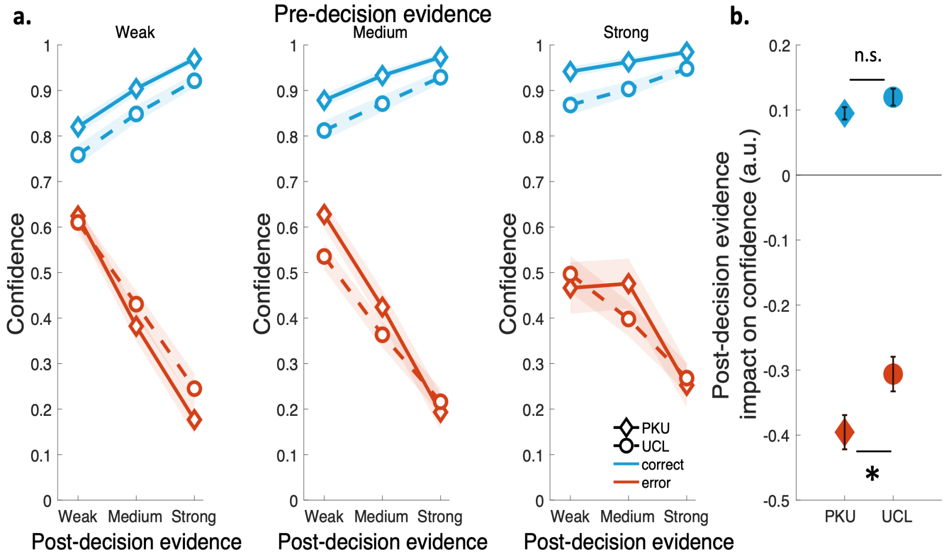


Figure 4.3: ***Behavioural results for Experiment 1. a.*** *Confidence as a function of post-decision evidence strength on error trials (red) and correct trials (blue) for each pre-decision evidence level. Shaded error bars represent group mean±SEM. N=39 at each site.* ***b.*** *Impact of post-decision evidence (PDE) on confidence indicated as standardized beta-coefficients from a hierarchical mixed-effect regression model on error trials (red) and correct trials (blue) at each site. Error bars represent group mean±SEM, *P<0.05.*

I next report how these effects interacted with site when also introducing the NYU dataset (setting PKU as a baseline in the regressions). In line with the site interactions described in chapter four, I find that the impact of post-decision evidence varied across sites (interaction post-decision evidence x site: $\chi^2(2) = 6.66$, $P = 0.04$). Contrasts show that this effect is mainly driven by a higher impact of post-decision evidence on confidence ratings in the PKU dataset than in the UCL dataset (contrast post-decision evidence PKU and UCL: $\beta = 0.06$, SE = 0.02, $t_{94.17} = 2.58$, $P = 0.01$). The contrast between PKU and NYU was in the same direction but did not reach significance (contrast post-decision evidence PKU and NYU: $\beta = 0.03$, SE = 0.02, $t_{100.74} = 1.02$, $P = 0.31$). As shown in **Figure 4.4a,b** the negative slope on error trials (red line) is steeper in PKU than in UCL or NYU. The three-way interaction between post-decision evidence, accuracy and site was not significant (interaction accuracy, post-decision evidence and site: $\chi^2(2) = 3.54$, $P = 0.17$). PKU participants were marginally more susceptible to post-decision evidence on error trials than NYU participants (β = -0.04, SE = 0.02, $t_{101.49} = -1.72$, $P = 0.09$; **Figure 4.4c**).

In addition to the hypothesized cultural differences in post-decision evidence processing, we also found cross-cultural differences in the impact of pre-decision evidence. The impact of pre-decision evidence varied across sites (interaction pre-decision evidence x site: $\chi^2(2) = 6.69$, $P = 0.04$). Contrasts reveal that pre-decision evidence had a lower impact on confidence in the PKU dataset than in the UCL dataset (contrast pre-decision evidence PKU and UCL: β = -0.05, SE = 0.02, $t_{92.37} = -2.56$, $P = 0.01$), the contrast between PKU and NYU is in the same direction but did not reach significance (contrast pre-decision evidence PKU and NYU: β = -0.03, SE = 0.02, $t_{98.84}$= -1.40, $P = 0.16$). In particular, this effect was restricted to error trials (3-way interaction accuracy, pre-decision evidence and site: $\chi^2(2)$= 9.18, P = 0.01). The impact of pre-decision evidence on error trials was lower in PKU than in UCL (contrast pre-decision evidence x accuracy: β = 0.05, SE = 0.02, $t_{95.22} = 2.80$, $P = 0.006$) and also lower in PKU than in NYU (contrast pre-decision evidence x accuracy : β = 0.05, SE = 0.02, $t_{101.96}$= 2.23, $P = 0.03$; **Figure 4.4c**).

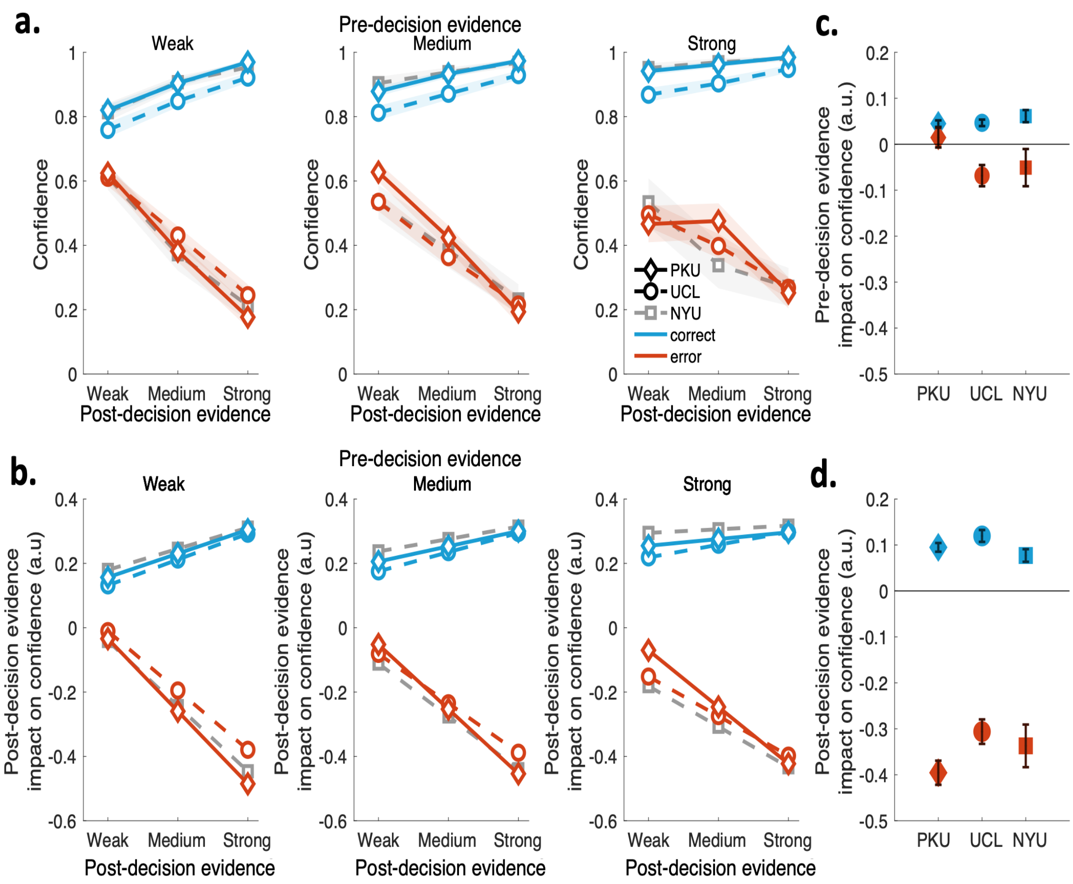


Figure 4.4: ***Behavioral results across PKU, UCL and NYU datasets. a.*** *Confidence as a function of post-decision evidence strength on error trials (red) and correct trials (blue) for each pre-decision evidence level. The NYU dataset is shown with dashed grey lines. Shaded error bars represent group mean ± SEM. N = 25 at NYU and N = 39 at UCL and PKU.* ***b.*** *Impact of pre-decision evidence level and post-decision evidence level on confidence as simulated from the beta-coefficients of the main hierarchical regression model reported in the main text (**Equation 4.2**).* ***c.*** *Impact of pre-decision evidence on confidence indicated as standardized beta-coefficients from a hierarchical regression model on error trials (red) and correct trials (blue) at each site. d, Impact of post-decision evidence on confidence indicated as standardized beta-coefficients from a hierarchical regression model on error trials (red) and correct trials (blue) at each site. Error bars represent group mean ± SEM.*

Finally, I asked whether a heightened sensitivity to post-decision evidence in the PKU group was also reflected in increased metacognitive efficiency (meta-d'/d' or Mratio). Note that the calculation of a metacognitive efficiency estimate in a post-decision evidence task departs from the usual usage of the meta-d' model in a task where sensory evidence is only available before a decision. However, fitting the model to the final confidence rating data provides a compact summary of the differential influence of various factors (including post-decision evidence) to metacognition across sites.

As such, metacognitive efficiency (meta-d'/d' or Mratio) was estimated using the HMeta-d toolbox (https://github.com/metacoglab/Hmeta-d) using Markov chain Monte Carlo (MCMC) sampling procedures within JAGS (http://mcmc-jags.sourceforge.net). Mratio for the PKU and UCL group was estimated separately across 30,000 samples after a burn-in of 1,000 samples distributed across three chains. To assess significance, I estimated the probability that the 95% highest density interval (HDI) of the estimated Mratio for PKU participants were higher than that of UCL participants: $P_{\theta}(PKU > UCL)$. I flag a ”significant” probability of $>$0.95 but also report the probability of a difference so readers can make their own decisions [Kruschke, 2010].

When I examined the HDI of the difference in the posterior distribution of estimates of metacognitive efficiency between sites (hierarchical estimation: 95% HDI [-0.07, 0.32], I found that 91% of the distribution was higher than zero ($P_{\theta}(PKU > UCL)$ = 0.91; **Figure 4.5**), indicating that metacognitive efficiency was higher among PKU than UCL participants.

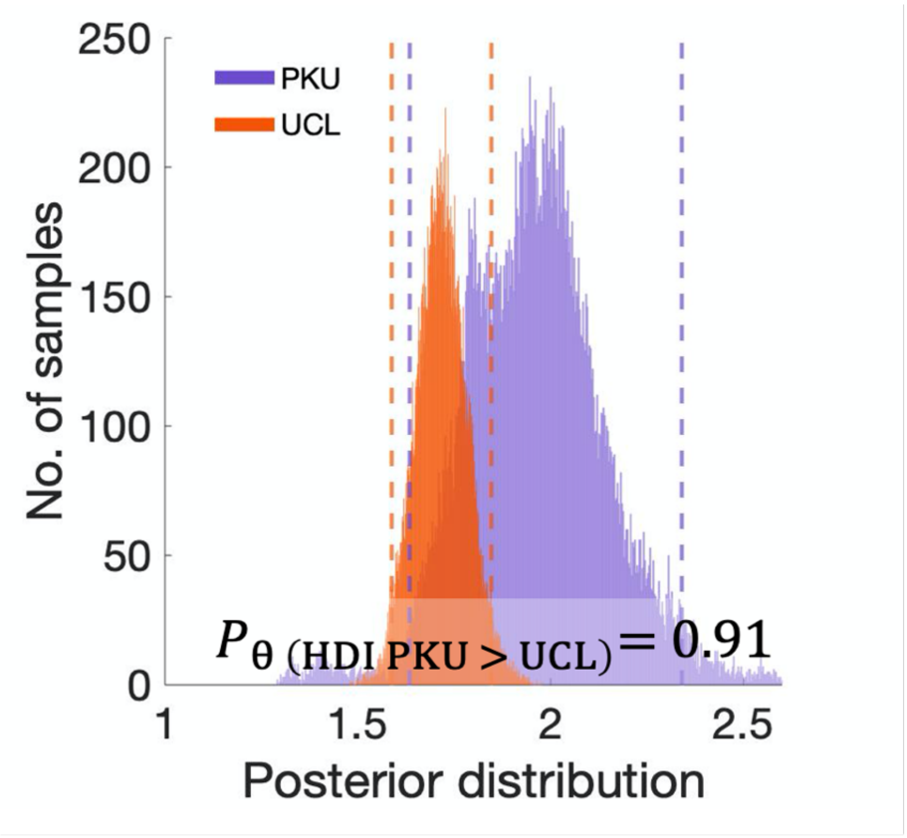


Figure 4.5: ***Metacognitive efficiency between sites.*** *Posterior distributions over group-level metacognitive efficiency for UCL and PKU participants separately. The dashed lines represent the 95% highest density intervals (HDI); $P\theta$ indicates the probability that the posterior samples of the PKU group are higher than the posterior samples of the UCL group.*

In summary, I found enhanced susceptibility to post-decision evidence in PKU participants compared with UCL participants, providing initial support for a heightened metacognitive evaluation of performance. Importantly, since first-order performance was matched between sites, these results support a hypothesis that metacognitive processes are liable to cultural influence.

4.4 Discussion

I here showed that participants with Chinese backgrounds were more susceptible to post-decision evidence than participants with British backgrounds. In particular, Chinese participants changed their minds more after errors than their British counterparts, consistent with enhanced metacognitive evaluation of performance facilitated by adaptive post-decision processing. Using a psychophysical task that enabled the separation of first-order and metacognitive processes in simple perceptual decisions, these results support a proposal that metacognition is sensitive to socio-cultural variation. Strikingly, these differences in confidence were found specifically on error trials, suggesting that cultural background may shape a metacognitive faculty to evaluate one's own performance.

These results are consistent with the recent theoretical proposal that explicit metacognition, the ability to self-evaluate one's perceptions, memories and decisions, is subject to cultural variation [Heyes et al., 2020]. The routes by which these differences emerge, and their stability over time, remains to be determined. One possibility is that the extent to which a culture places emphasis on the group over the individual may make it more likely that the skills needed to question and doubt one's beliefs and decisions are culturally inherited. For instance, in more collectivist societies there may be greater advantages to be gained by honing the sharing and communication of accurate confidence estimates [Bang et al., 2017, Mahmoodi et al., 2015]. In contrast, in more individualistic societies, cultivating distorted metacognition for one's own ends (e.g., an overconfident style) may be prioritized. It also remains unclear as to what aspects of self-evaluative processing are affected by culture. In previous studies using related tasks within cultures, a distinction has been drawn between brain areas that are sensitive to post-decision evidence (in posterior medial frontal cortex) and those in more anterior frontal regions that mediate a mapping between private and public aspects of confidence [Bang et al., 2017, Bang et al., 2014, Fleming et al., 2018, Gherman and Philiastides, 2018]. Either or both of these levels of processing may plausibly be affected by culture and, at both an individual and group level, contribute to the current results.

This study aimed at a robust and replicated assessment—using new, sensitive and specific methods that provide an in-depth analysis of individuals' metacognitive processes–to compare two closely-matched samples drawn from distinct cultural milieus (for which *a priori* evidence suggested cross-cultural differences) and so provide evidence for or against an important hypothesis regarding human metacognition. It is important to note that neither China's or any other state or region's culture is monolithic, and these samples are by no means representative of all Chinese or UK citizens. Instead, I chose to investigate two well matched subgroups. The strengths of such a tightly controlled, robust and replicated approach to explore a specific hypothesis can be complemented by future work using other approaches, which can, for example, look across broader groups of samples drawn from other ages, different socio-economic backgrounds, different levels of education (including adaptations to semi-literate populations) and other regions (within Northern Europe, within China and globally). Combining diverse types of study—both tightly controlled studies and those testing greater generalizability [Tiokhin et al., 2021]—will likely provide greater advances in understanding of human cognition and its cultural contributions than either types of study alone.

Another limitation is that I did not explicitly account for motor errors in my task. It could be that motor errors lead to distinct types of post-decision evidence processing. As first-order performance was matched, such that the overall number of errors similar between cultures, I consider it unlikely that motor errors will have affected the current results. An open question is whether similar results would be obtained using a confidence task without post-decision evidence. Previous work has found that the effect of post-decision evidence on error trials is strongly correlated with metacognitive efficiency in a task that does not involve presentation of post-decision evidence [Rollwage et al., 2018]. I therefore believe that similar cultural differences would have been found even in the absence of a post-decision evidence manipulation—but because I did not measure confidence immediately after the decision here, this remains an open question for future work.

In summary, I demonstrated that populations with Chinese backgrounds demonstrate heightened metacognitive evaluations of performance in comparison with populations with British or American backgrounds. These differences manifested in boosts to post-decisional processing following error trials, in the absence of differences in first-order performance. These results provide initial evidence that socio-cultural background can shape the tendency to evaluate and reflect on previous decisions.

Chapter 5

Social contributions to metacognition

5.1 Introduction

The previous chapter outlined the principles of the cultural origins hypothesis which suggests that aspects of metacognition may be susceptible to cultural norms of collaboration. Environments that emphasize shared over individual goals, may promote forms of cognition that facilitate successful collaboration, such as openness to discussing the mental states of self and other [Cleeremans et al., 2020, Heyes et al., 2020, Heyes and Frith, 2014]. I showed that the computational process that gives rise to subjective confidence judgments is different between Chinese and UK or US populations. Specifically, I compared how a Chinese and British sample matched for occupation, age, gender, income and IQ, integrated new evidence to evaluate the probability that a previous choice was correct. In line with the Cultural origin hypothesis, I found that the sample with a Chinese background showed a heightened sensitivity to new evidence—which led them to recognize and correct their errors more than UK and US samples.

The ability to consciously evaluate and interpret mental states, such as the ability to recognize one's own mistakes, may be culturally acquired similar to the ability to read books [Heyes, 2018]. When a child learns how to read their new skill is not restricted to reading just one book but can be applied to read a variety of different books. Similarly, the ability to process new evidence and recognize one's errors should be acquired in a domain-general manner: as a global mechanism that can be applied as the ability to understand one's own metacognitive processes (e.g., in giving reliable advice to others) as well as other people's mental states (mentalizing; e.g., in knowing when other people's social advice is reliable; [van der Plas et al., 2019]). This putative domain generality of metacognition is a recent topic of discussion and some evidence suggests that metacognition in processing new evidence indeed works similarly across various distinct types of information (e.g., perceptual and numerical evidence are processed similarly; [Bronfman et al., 2015, Talluri et al., 2018]). Later research has found that being faced with an opinion that challenges a personally held believe (social evidence) and privately being presented with more of the same evidence (non-social evidence) follows a similar mechanism [Behrens et al., 2008].

On these advice-taking paradigms, participants make a first decision and are then presented with the opinion of an 'adviser' [De Martino et al., 2017, Campbell-Meiklejohn et al., 2017, Campbell-Meiklejohn et al., 2010] [Behrens et al., 2008, Gomez-Beldarrain et al., 2004, Sniezek and Van Swol, 2001]. Interestingly, and in line with the finding that more reliable evidence elicits more changes of mind [Bronfman et al., 2015, Fleming et al., 2018, Talluri et al., 2018], the reliability of the advice is a crucial determinant of how much it engenders a change in peoples' beliefs. In advice-taking settings, this reliability of the advice can be communicated in the form of the confidence of the adviser [Campbell-Meiklejohn et al., 2017, Campbell-Meiklejohn et al., 2010, Gomez-Beldarrain et al., 2004, Sniezek and Van Swol, 2001], as judgments made with higher confidence are typically more likely to be correct.

In addition to the unavoidable potential for a mistake in perceptual decision making, social advice has an additional cause for error, as some people are intentionally deceptive or unreliable. In other words, efficient advice-taking involves assessing not only the probability that oneself is correct (via metacognition) but also the reliability of the adviser (via mentalizing; [Burke et al., 2010, De Martino et al., 2017, Pescetelli and Yeung, 2021, Harvey and Fischer, 1997]). Advisors' expressions of certainty are typically a useful source of information about the advisers' reliability, as people who say that they are confident are usually also more likely to be correct [Campbell-Meiklejohn et al., 2017, Campbell-Meiklejohn et al., 2010, Gomez-Beldarrain et al., 2004]. This process is complicated when the fidelity of the advisers' confidence ratings is not representative of their accuracy (the adviser's metacognitive ability). Put differently, it is often sensible to take advisers' certainty estimates with a pinch of salt [Bahrami et al., 2010, Bahrami et al., 2012, Bang et al., 2017], and learn, over the course of repeated interactions, which advisers' confidence estimates are more reliable than others [Hertz et al., 2017, Pescetelli and Yeung, 2021].

It could be that both assessing one's own need for new evidence follows a similar set of mental shortcuts used to evaluate whether another person's advice is reliable, as would be expected under the Mentalizing is prior hypothesis [Carruthers, 2009, Gazzaniga, 1995, Gazzaniga, 2000, Gopnik, 1993, Wegner, 2002, Wilson, 2002]. If this idea is correct, collaborative environments should facilitate the process of understanding when more evidence is needed (metacognition) to a similar extent as an understanding which types of advice are more reliable (mentalizing). In addition, if the Cultural Origins Hypothesis is domain-general, I would expect advice-taking and metacognitive processing to be similarly boosted in Chinese populations compared with UK populations. Instead, if metacognition and mentalizing develop from distinct sources of input, one would expect post-decision evidence processing and advice-taking to be distinctly malleable to cultural differences in norms of collaboration.

One caveat with studying whether a domain-general tendency for advice-taking is different for people with distinct cultural backgrounds, is that there are reputational [Bhaskar and Thomas, 2019, Tenney et al., 2019] and evolutionary

[Johnson and Fowler, 2011] advantages for maintaining a higher confidence than would be warranted on the basis of true accuracy. Given that social and reputational benefits are likely to vary across distinct social settings, overall confidence bias may also vary across distinct cultures (which has indeed been shown to be the case in previous work: [Yates et al., 1998, Yates et al., 1989]). In other words, there may be cultural differences in what level of confidence is considered “high”–which may have been a confound in earlier investigations of advice-taking across cultures.

I here circumvent this problem by experimentally calibrating the social advice to participants’ own responses. Just as in the previous experiment, I present three levels of post-decision evidence strength (weak, medium, strong) that are crossed with three levels of pre-decision evidence (weak, medium, strong) towards a three (pre-decision evidence strength) x three (post-decision evidence strength) within-subject conditions. However importantly, on a randomly selected half of the trials, post-decision evidence consisted of the confidence estimation of a previous participant (‘adviser’) as social post-decision evidence. In particular, the social advice was obtained from a generative model that had the same perceptual sensitivity as the participant. This model allowed me to control the informativeness of social and non-social evidence and ensure that the confidence levels of the advisers followed the three evidence strength conditions. From advice-taking trials, I was able to define metacognitive efficiency as the ability to rely more on advice when oneself is wrong (vs. correct); and mentalizing efficiency as the ability to rely more on advice when the adviser is right (vs. wrong). This method allows me to ensure that the used metric of metacognition and mentalizing are unconfounded by differences in first order cognition, such as the adviser’s choice accuracy or confidence bias.

As in the previous chapter, I again recruited two new samples of Chinese and British cultural backgrounds, that were carefully matched for occupation, age, income, IQ and gender, and replicate the metacognitive advantage of Chinese participants over British participants is also obtained when the new evidence consists of social advice. In particular, Chinese participants have boosts in post-decisional processing following error trials, irrespective of whether the evidence is of a social or perceptual type. Moreover, I find that this effect is restricted to trials on which the participant is wrong, and the adviser correct—suggesting a cultural benefit in advice taking for persons that grew up in cultures where collaborations and groups are placed before the individual.

5.2 Methods

5.2.1 Participants

I recruited two new samples of Chinese and British cultural backgrounds, that were carefully matched as in the previous chapter. A minimum sample size of N = 53 at each site was defined by an a priori power calculation of the t-test between the impact of post-decision evidence on confidence in PKU and UCL in the previous chapter (power = 80%, $P = 0.05$, Cohen's d = 0.54). This power calculation provides a simple, relatively assumption-free estimate of effect size for our key contrast of interest. Four participants were excluded from the PKU dataset: one participant performed below an a priori accuracy cut-off of 60%; two participants' calibration data was unusable, and one participant violated transitivity in performance (i.e., average performance was lower in the medium evidence condition than in the weak evidence condition). Two participants were excluded from the UCL dataset: one participant did not believe the social manipulation and never followed the advice (see **Experimental paradigm**), the other participant violated transitivity. All reported site differences of post-decision evidence on confidence remained significant after I re-introduced these excluded participants. All participants had normal or corrected-to-normal vision and no history of neurological or psychiatric illness. The study was approved by the University College London Ethics Committee (1260/003) and by the Ethics Committee of School of Psychological and Cognitive Science at Peking University. All participants gave written informed consent before taking part in the experiment. The total duration of the task was approximately 40-50 minutes and participants were instructed to take six self-paced breaks at specific moments in the task.

5.2.2 Experimental paradigm

I adapted the task used in the previous chapter. As in the original task, participants were asked to judge the direction of moving dots (pre-decision evidence) with varying evidence strength (weak, medium or strong). I made a number of changes to the original paradigm. Confidence ratings were made on a confidence scale that ranged from 100% confidence in the left direction to 100% confidence in the right direction (100%, 80%, 60% left and 60%, 80%, 100% right). Participants were asked to rate their confidence on this scale because, on a randomly selected half of the trials, the same scale was used to display the confidence estimation of a previous participant ('adviser') as social post-decision evidence. On the other half of the trials, post-decision evidence was a second RDK stimulus with dots moving in the same direction as pre-decision evidence but with variable evidence strength (weak, medium, strong). Social post-decision evidence was presented below a silhouette with a unique, uninformative background color. Participants were told that, because of the calibration procedure, the performance of the advisers was similar to theirs. All but one of my 106 participants across both sites indicated to have believed the social manipulation during my extensive debriefing. In reality, the social advice was obtained from a computational model that made decisions with the same perceptual sensitivity level as the participant. This manipulation allowed me to keep the informativeness of post-decision evidence equal across conditions (social, perceptual) and manipulate the confidence levels of the adviser as a function of three evidence strength levels (with more confident advisers following stronger evidence; **Figure 5.1**).

Adviser's responses (a_{adv}) were simulated under a signal detection theoretic model. We computed the perceptual sensitivity levels (d') that an adviser who had experienced the same calibration procedure as subjects should show for each level of simulated evidence strength. This was obtained by transforming the target probability correct values (P_{adv}) used in calibration:

$$\begin{gathered} P_{adv} = [0.6,\ 0.75,\ 0.9] \\ d' = 2 * norminv\,(P_{adv}) = \ [0.507,\ 1.3491,\ 2.563] \end{gathered} \tag{5.1}$$

From d' we could calculate samples of evidence experienced by the adviser on each trial (x_{dir}), sampled from a normal distribution ($\sim$ N) with mean determined by the perceptual sensitivity on a given trial (s, [weak: 1, medium: 2, strong: 3]), sign dependent on the true direction of the dots (dir, indicated as [left: -1, right: 1]) and a standard deviation of 1:

$$x_{dir} \sim N\left(dir * \tfrac{d\prime(s)}{2},\ 1\right) \tag{5.2}$$

The adviser reported rightward movement (a = 1) if exceeded an internal decision criterion which we assumed to be unbiased [m = 0]:

$$\begin{gathered} if\,(x_{dir} > m) \\ a_{adv} = 1 \\ else \\ a_{adv} = \ -1 \end{gathered} \tag{5.3}$$

In addition to generating the choices of the adviser (a_{adv}), we used the same signal detection theory model to generate trial-by-trial adviser confidence levels. Due to an error in this model that misspecified the mean and variance during inference, advisers were generally less confident than most participants. Despite this general tendency towards under-confidence, adviser confidence levels mimicked key features of human confidence levels: advisers were generally more confident about correct decisions (**Figure 5.1a**) and less confident about wrong decisions (**Figure 5.1b**). Furthermore, adviser confidence on correct trials was lowest in the weak post-decision evidence condition (57%, SE = 0.01), higher in the medium post-decision evidence condition (67%, SE = 0.01) and highest in the strong post-decision evidence condition (76%, SE = 0.02). To avoid relying on the model when analyzing data, we decided to bin adviser confidence into three levels based on the 33% inter-quartile cumulative distribution and entered this as social post-decision evidence [-0.5: weak, 0: medium, 0.5: strong] in all regression analyses.

Together, this full-factorial design crossed three (pre-decision evidence strength) x three (post-decision evidence strength) x two (social, perceptual post-decision evidence type) within-subject conditions.

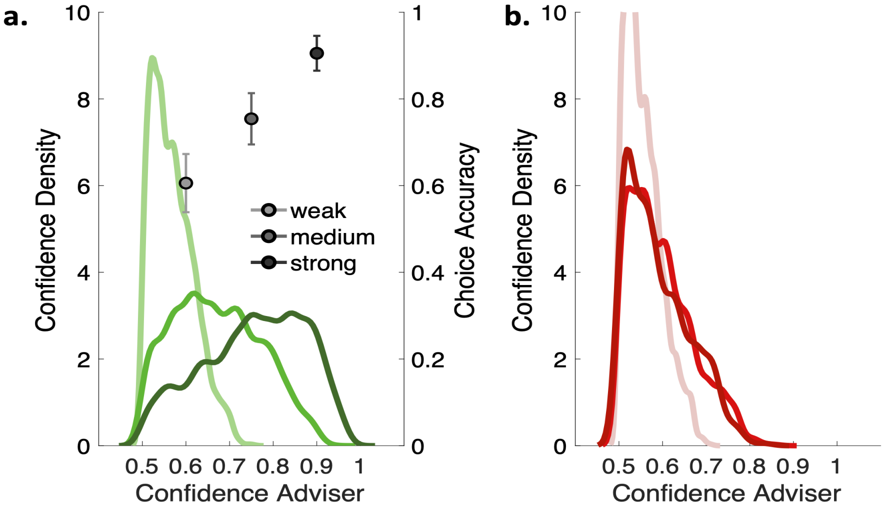


Figure 5.1: ***Confidence levels of advisers.*** ***a.*** *The probability density distributions of adviser confidence levels on correct trials. The left y-axis represents the probability density estimate along the three post-decision evidence levels [weak, medium, strong]. The right y-axis represents advisers' choice accuracy for each level of post-decision evidence. Error bars represent group mean ± SEM.* ***b.*** *The probability density distributions of adviser confidence levels on error trials. The y-axis represents the probability density estimate along three post-decision evidence levels (weak, medium, strong).*

5.2.3 Additional measures

In addition to the three questionnaires administered in the previous chapter: the Self-Construal Scale [Singelis, 1994], Cattell Culture Free Intelligence Quotient [Cattell, 1943] and the Analysis Holism Scale [Choi et al., 2003]. I also obtained participant's responses on the Beck Cognitive Insight Scale (BCIS; [Beck et al., 2004]). This scale was originally developed to measure insight into symptoms within clinical populations but has also been used in non-clinical settings [Fleming and Dolan, 2012]. On the BCIS, participants indicated their agreement with statements about the recognition that experienced reality may be different from the objective truth. A person's tendency to reflect on their inner experiences is captured with the 'self-reflectiveness' subscale; and their ability to critically reconsider inner experiences based on counterevidence is captured with the 'self-confidence' subscale [Beck et al., 2004]. Participants rated their agreement with fifteen items on a scale from 0 ('do not agree at all') to 3 ('agree completely'), from which I computed a main composite score following a standard coding procedure. My reported scores on the BCIS were comparable with the scores of a control group in a large-scale clinical study conducted in India [Jacob et al., 2019], proposed that BCIS may differ between collectivist versus individualist cultures but did not test this empirically.

I was interested in knowing how insight would relate to differences in post-decision evidence processing on the main task and whether, in light of the cultural variation hypothesis, I would find cross-cultural differences on the BCIS (**Table 5.1**).

5.2.4 Statistics

Statistical inference was conducted similarly to analysis of **chapter four**. As confidence estimates were given on a different scale in the previous chapter, I first converted confidence in left and right (confdir) to confidence in the chosen direction [certainly wrong: 0, certainly correct: 1], by subtracting $conf_{dir}$ from 1 when the chosen direction was left (a = -1), following:

$$\begin{aligned} if : a == -1 \\ conf = 1 - conf_{dir} \\ else : conf = conf_{dir} \end{aligned} \tag{5.4}$$

To index the strength of social post-decision evidence while ignoring the direction of the advice, I transformed adviser confidence ($conf_{adv}$) on a scale from 100% left to 100% right. I recoded this variable as ranging from 0-1, such that values < 0.5 indicated greater adviser confidence in leftward motion and values > 0.5 indicated greater adviser confidence in rightward motion. I then transformed this signed confidence variable to an unsigned confidence variable ranging from 0.5 to 1, as follows:

$$\begin{aligned} ifconf_{adv} < 0.5 \\ conf_{adv} = 1 - conf_{adv} \end{aligned} \tag{5.5}$$

Then I binned adviser confidence into three equal quantiles representing the lowest, middle and highest 33% confidence ratings (confadv) to create 3 levels of social post-decision evidence [weak: -0.5, medium: 0, strong: 0.5], which I used instead of ‘post’ in Equation 3.2 in the previous chapter.

Each individual’s beta coefficient for the main effect of perceptual and social post-decision evidence (derived from Equation 3.3 in the previous chapter) were entered into a robust correlation using the Matlab robust correlation toolbox [Pernet et al., 2013].

5.3 Results

In order to replicate and extend the results from the previous chapter I recruited two new samples of N = 53 PKU participants (25 females, M_{age} = 21.91 (SE = 0.46) and

Table 5.1: ***Demographical and trait differences between PKU and UCL datasets.*** *Group mean ± standard error from the mean. Income in pounds (£) is given relative to the purchase power parity (PPP) ratio between UK and China at the time of recruitment (ratio 1:1.71). Mean composite score ± standard error from the mean is given for the CFIT (Cattell Culture-Free Intelligence Test), AHS (Analysis-Holism scale), ind (independent) and int (interdependent) SCS (Self-Construal Scale), main composite BCIS (Beck Cognitive Insight Scale), and the self-reflectiveness (sr) and self-confidence (sc) subscales of the BCIS. *$P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ for the independent samples t-test contrast between sites (assuming equal variance), bold values indicate a significant group difference.*

Trait	**PKU**	**UCL**
age	21.91 ± 0.46	22.49 ± 0.41
% female	47%	55%
income(£)/PPP	41,373 ± 5,454	56,989 ± 13,767
CFIQ	99.21 ± 1.41	102.00 ± 1.46
AHS	**5.05 ± 0.07**	**4.73 ± 0.06**
SCS-ind	4.86 ± 0.09	4.87 ± 0.12
SCS-int	4.74 ± 0.12	4.76 ± 0.09
BCIS	**10.17 ± 0.60**	**6.28 ± 0.73** ***
BCIS-sr	**25.19 ± 0.42**	**13.62 ± 0.60** ***
BCIS-sc	**15.02 ± 0.35**	**7.34 ± 0.50** ***

N = 53 UCL participants (29 females, M_{age} = 22.49 (SE = 0.41), again with similar age (t_{104} = -0.95, 95% CI = [-1.81, 0.64], P = 0.34), gender (M_{PKU} = 49%, M_{UCL} = 51%, t_{104} = -0.78, 95% CI = [-0.27, 0.12], P = 0.44), Culture Free Intelligence Quotient (M_{PKU} = 99.21 (SE = 1.41), M_{UCL} = 102.00 (SE = 1.46), t_{102} = -1.37, 95% CI = [-6.82, 1.24], P = 0.17) and annual family income (M_{PKU} = £41,373.58 (SE = 5,454.69) and UCL (M_{UCL} = £56,988.89 (SE = 13,766.63),$t_1$02 = -1.05, 95% CI = [-45060, 13830], P = 0.30) were recruited. In light of the findings of enhanced self-evaluation in PKU participants in the previous chapter, I hypothesized that PKU participants would report having greater insight than UCL participants. This hypothesis was confirmed by the questionnaire data, with PKU participants having higher average BCIS scores than UCL participants (M_{PKU} = 40.26 (SE = 0.49); M_{UCL} = 20.96 (SE = 0.82), independent samples t-test, t_{104} = 20.08, 95% CI = [17.39, 21.21], $P < 2.2\text{e-}16$; see **Table 5.1**).

Participants again made a binary perceptual discrimination (left versus right random dot motion) based on pre-decision evidence of varying strength (weak, medium or strong). Half of the trials were similar as the task used in chapter four (using perceptual post-decision evidence). In the other half of trials, perceptual post-decision evidence was replaced by the confidence and direction judgment provided by an anonymous previous participant ('adviser'). This manipulation allowed me to assess whether cultural differences in post-decision processing would generalize across different domains (perceptual, social). In practice, I generated adviser choices from a model that mimicked the perceptual sensitivity of the participant. The stimulus that I presented to the simulated adviser was that trial's perceptual post-decision evidence level, i.e., the evidence strength that would have been presented to the participant in the equivalent perceptual condition (with the same dot direction as the participant's pre-decision evidence yet with potentially variable strength). As a result of this, adviser accuracy and confidence levels were contingent on the perceptual post-decision evidence strength on any particular trial, which was counterbalanced with respect to the pre-decision evidence strength just as for the perceptual condition. Participants were paired with a new adviser on every trial and were told that all advisers had the same accuracy in detecting the motion direction as themselves due to completion of an identical calibration procedure. One participant reported not to believe the social manipulation and was excluded from further analyses (see **Methods**).

I defined social post-decision evidence strength as the adviser's confidence rating binned into three levels (low, medium, high), creating a fully factorial 3 (pre-decision evidence strength) x 3 (post-decision evidence strength) x 2 (post-decision evidence type) design (**Figure 5.2a**). Using a hierarchical logistic regression on trial-by-trial accuracy in Experiment 2, I confirmed that choice accuracy was higher when participants had seen stronger pre-decision evidence (main effect pre-decision evidence: $\chi^2(1)$ = 484.85, $P <$ 2e-16, β = 2.97 (SE = 0.14). As per the calibration procedure, this effect did not interact with site (no interaction-effect pre-decision evidence and site: $\chi^2(1)$ = 0.003, P = 0.96, β = -0.01 (SE = 0.19) nor post-decision evidence type (no interaction-effect pre-decision evidence level and post-decision evidence type: $\chi^2(1) = 0.0003$, $P = 0.99$, $\beta = -0.01$; SE = 0.19).

As in the previous chapter, I ensured that first-order performance was matched across participants and across both post-decision evidence types (**Figure 5.2b**). I also did not find a difference in average confidence across sites (M_{PKU} = 82% (SE = 0.01), M_{UCL} = 79% (SE = 0.01), independent samples t-test, t_104 = 1.64, 95% CI = [-0.01, 0.06], P = 0.10).

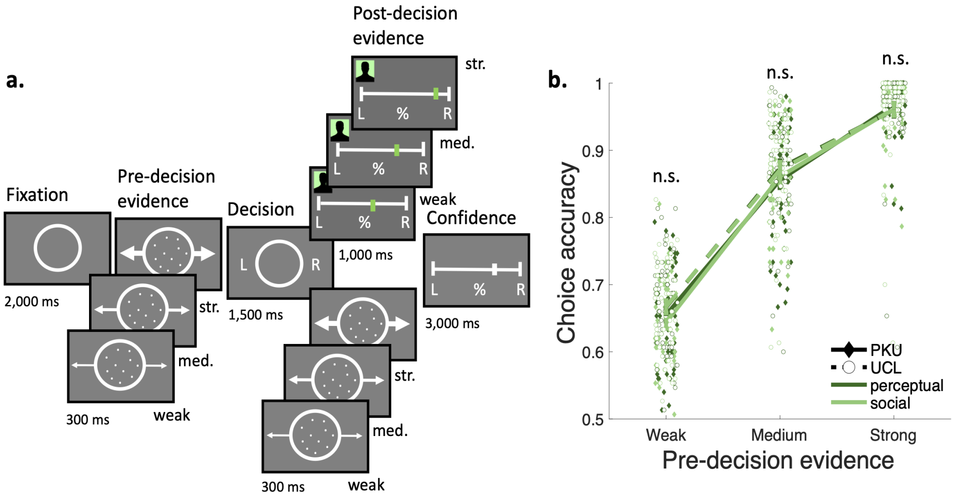


Figure 5.2: ***Task design and first-order performance. a.*** *Participants were asked to make judgments about the direction (left, right) of random dot motion stimuli. Afterwards participants were either shown perceptual post-decision evidence or with what an anonymous 'adviser' had decided on the same trial (social post-decision evidence, which was generated from a computational model). At the end of each trial, participants were asked to rate their confidence that the initial decision was correct on a scale from 100% left-stimulus to 100% right-stimulus.* ***b.*** *Choice accuracy was matched between sites (n.s.) and higher following stronger pre-decision evidence levels (P<0.001, N=53 at each site). Error bars represent group mean±SEM.*

In the perceptual condition, I replicated the findings from the previous chapter that PKU participants, in comparison with UCL participants, have heightened metacognitive evaluation when processing post-decision evidence. Specifically, perceptual post-decision evidence had a higher impact on confidence in the PKU dataset than in the UCL dataset (hierarchical linear regression, interaction perceptual post-decision evidence x site: $\chi^2(1) = 10.39$, $P = 0.001$, β= 0.06 (SE = 0.02);

Figure 5.3a). This effect was again most evident on error trials, which led to a significant three-way interaction (hierarchical linear regression, interaction perceptual post-decision evidence x accuracy x site: $\chi^2(1)$ = 7.07, P = 0.008, β= -0.05 (SE = 0.02).

I next asked whether a cultural difference in metacognition would generalize to a situation in which post-decision evidence is presented as social advice. In the social condition, I calculated how often participants changed their mind towards the direction suggested by the adviser on trials in which the participant and adviser disagreed. This tendency to change one's mind and comply with the adviser was higher in PKU participants than in UCL participants (M_{PKU} = 17.9%, M_{UCL} = 12.6%, independent samples t-test, t_{104} = 2.21, 95% CI = [0.005, 0.10], P = 0.03). In keeping with a metacognitive advantage in PKU participants, this effect was restricted to trials on which the participant was wrong (and accordingly, the adviser correct; M_{PKU} = 33.8%, M_{UCL} = 24.1%, independent samples t-test, t_{104} = 2.59, 95% CI = [0.02, 0.17], P = 0.01), and was not seen on trials in which the participant was correct (and the adviser wrong; M_{PKU} = 8.3%, M_{UCL} = 6.5%, independent samples t-test, t_{104} = 0.92, 95% CI = [-0.02, 0.06], P = 0.36). This result suggests that the cross-cultural asymmetries in post-decision processing identified using perceptual stimuli generalize to cases in which new evidence is presented as social advice.

To further examine the drivers of cross-cultural differences in advice-taking, I inferred the impact (beta coefficient) of adviser confidence [low, medium, high] on participants' confidence levels using a hierarchical mixed-effect model. Similar to the cross-cultural differences in perceptual post-decision evidence processing reported in chapter four and five, advice had a greater impact on the confidence ratings of PKU participants compared to UCL participants (hierarchical linear regression, interaction between social post-decision evidence x site: $\chi^2(1)$ = 8.38, P = 0.004, β= 0.04 (SE = 0.02). As expected from the previous analyses, this asymmetry in the impact of adviser confidence was most evident on trials where the participant made an error (hierarchical linear regression, interaction social post-decision evidence x initial choice accuracy x site:$\chi^2(1)$= 10.56, P = 0.001, β= -0.05 (SE = 0.02); **Figure 5.3a**), consistent with a hypothesis of cultural differences in metacognitive evaluation of performance.

At both sites, social post-decision evidence had a lower impact on confidence than perceptual post-decision evidence (hierarchical linear regression, interaction evidence type x post-decision evidence strength: $\chi^2(1)$ = 77.34, $P <$ 2.2e-16, β = 0.06 (SE = 0.007). However, an enhanced susceptibility to post-decision evidence in PKU compared with UCL participants was found irrespective of whether the evidence was social or perceptual (no three-way interaction between evidence type, post-decision evidence and site: $\chi^2(1)$ = 3.35, P = 0.07, β = -0.02 (SE = 0.01).

The similar manner in which social and perceptual post-decision evidence was processed suggests a domain-general component of post-decision evidence processing, which is in line with previous work [Rouault et al., 2018, Carpenter et al., 2019a]. In line with the pattern of confidence reports obtained in the perceptual version of the task, participants across both sites reported higher confidence after receiving more confident confirming advice and lower confidence after receiving more confident disconfirming advice (hierarchical linear regression, interaction-effect of social post-decision evidence and accuracy: $\chi^2(1)$ = 93.18, P = 2.2e-16, β= 0.08 (SE = 0.01; **Supplementary Material 2.3**). To further investigate this putative domain-generality, I next asked whether the impact of perceptual and social post-decision evidence was similar for any given individual. **Figure 5.3b**) shows that this was the case: the impact of these two evidence types were positively correlated among both PKU participants (robust correlation, r = 0.45, 95% CI = [0.19, 0.64], P = 0.0006) and UCL participants (robust correlation, r = 0.39, 95% CI = [0.13, 0.64], P = 0.004), suggesting that participants who are more likely to integrate new perceptual evidence to update their confidence are also more likely to make use of social advice.

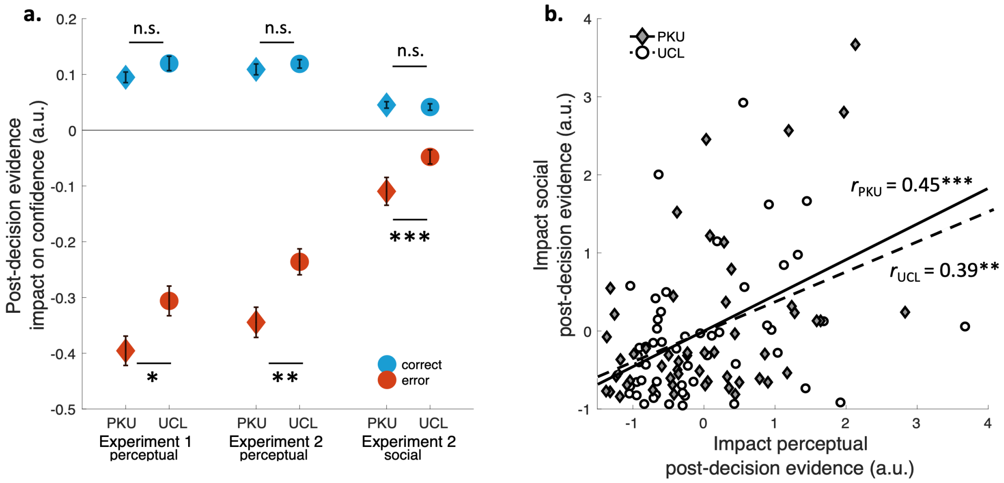


Figure 5.3: ***Cultural differences in changes of confidence.*** ***(a)*** *Impact of perceptual and social post-decision evidence on confidence on error trials (red) and correct trials (blue) across sites and experiments. The coefficients from chapter four (**Figure 4.3b**) are replotted for comparison.* ***(b)*** *Standardized beta-coefficients for the impact of perceptual and social post-decision evidence on confidence for each participant from a hierarchical mixed-effect regression model standardized within each site. Error bars represent the group means±SEM, ***P<0.001, **P<0.01 and *P<0.05.*

A key difference between social and perceptual evidence is that perceptual post-decision evidence is always in the correct direction, whereas the advisers could sometimes be wrong. As a result, there were four possible trial scenarios in the social condition: (1) The participant was correct, and the adviser agreed ('good' agreement); (2) The participant as wrong and the adviser disagreed ('good' disagreement); (3) The participant was correct, yet the adviser disagreed ('bad' disagreement); (4) The participant was wrong, yet the adviser agreed ('bad' agreement).

To facilitate exploratory analyses of social post-decision evidence, I here transformed participants' confidence in the chosen direction to confidence in the objectively correct direction (ranging from higher confidence in the incorrect direction to higher confidence in the correct direction) as the dependent variable in an extended hierarchical regression model (**Equation 5.3**), to which I introduced agreement between the participant and adviser [disagree: -1, agree: 1] as an additional predictor variable. Confidence in the objectively correct direction was higher on disagree trials than on agree trials (main effect of agreement: $\chi^2(1)$ = 24.54, *P* = 7.28e-07, β = -0.07 (SE = 0.01). This effect is explained by a two-way interaction with accuracy, indicating that 'good' agreement increased participants' confidence in the objectively correct direction, yet to a smaller extent than 'bad' agreement decreased participants' confidence in the objectively correct direction (interaction effect of agreement x accuracy: $\chi^2(1)$ = 82.74, *P* = 2.2e-6, β = 0.26 (SE = 0.03). When I allow the predictors to interact with site, I find that more confidently disagreeing advisers had a more pronounced impact on PKU participants than UCL participants (interaction agreement x social post-decision evidence level x site: $\chi^2(1)$ = 5.53, *P* = 0.02, β = 0.06 (SE = 0.03). PKU participants were especially more susceptible to strongly disagreeing advisers when their initial decision was, in fact, wrong (interaction accuracy x agreement x confidence adviser x site: $\chi^2(1)$ = 5.55, *P* = 0.02, β = -0.11 (SE = 0.05). This effect is shown in **Figure 5.4a** with the consistently steeper upwards sloping dark red lines in PKU than in UCL; and in **Figure 5.4b** with heightened impact of 'good' disagreement in PKU than in UCL (compare the dark red dots across sites), but a similar impact of 'bad' disagreement across sites (compare the light blue dots across sites). In sum, these results show that PKU participants were more influenced by social post-decision evidence than UCL participants, but only when the advice was useful.

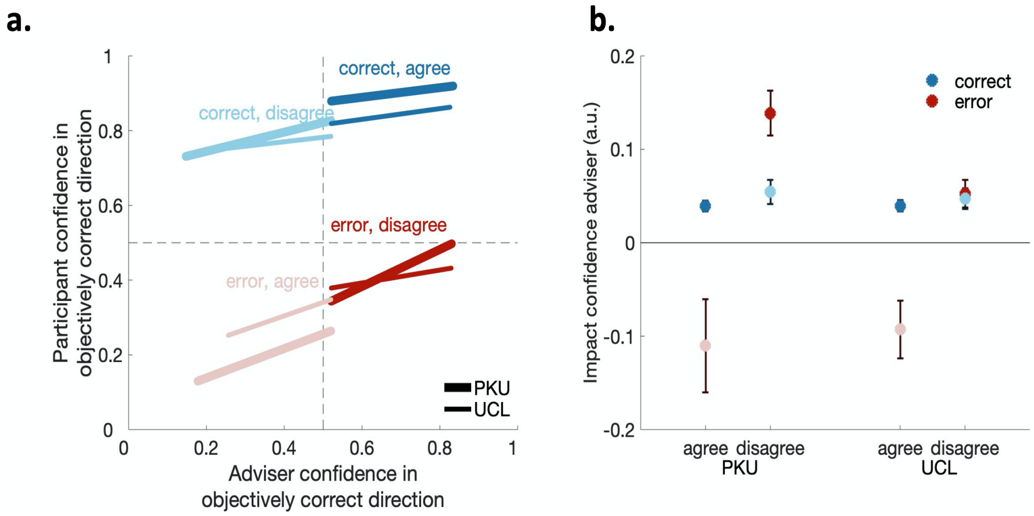


Figure 5.4: ***Cross-cultural differences in social advice-taking. a.*** *Linear fit of the association between the participant's confidence in the objectively correct direction and the advisers' binned confidence level in the objectively correct direction on correct (blue) and error trials (red) across sites, and as a function of adviser accuracy (good advice in darker colours, bad advice in lighter colors).* ***b.*** *The impact of advisers' binned confidence level in the objectively correct direction on the participant's confidence in the objectively correct direction on error (red) and correct trials (blue) across sites, and as a function of adviser accuracy (good advice in dark-er colors, bad advice in lighter colors), indicated as standardized beta-coefficients from a hierarchical mixed-effect regression model. As expected, the impact coefficients of good advice resemble the impact coefficients of perceptual post-decision evidence (as shown in* ***Figure 5.3a****).*

These results suggest that metacognitive efficiency was higher in PKU participants than in UCL participants. To test for such a difference, I next estimated metacognitive efficiency using the HMeta-d toolbox for each group separately. In line with the findings of Experiment 1, PKU participants had higher metacognitive efficiency than UCL participants in the perceptual condition (hierarchical estimation: HDI: [-0.051, 0.164], $P_{\theta\ (PKU\ >\ UCL)}\ = 0.853$; **Figure 5.5a**), albeit not significantly so. I conducted a similar analysis for social condition, where I again found that metacognitive efficiency was higher in PKU than UCL participants (hierarchical estimation: HDI: [-0.05, 0.16], $P_{\theta\ (PKU\ >\ UCL)}\ = 0.98$, **Figure 5.5b**).

In a final analysis I explored whether individual differences on the questionnaires were associated with beta coefficients on the main task. I computed each participant's interaction coefficient between accuracy x social/perceptual post-decision

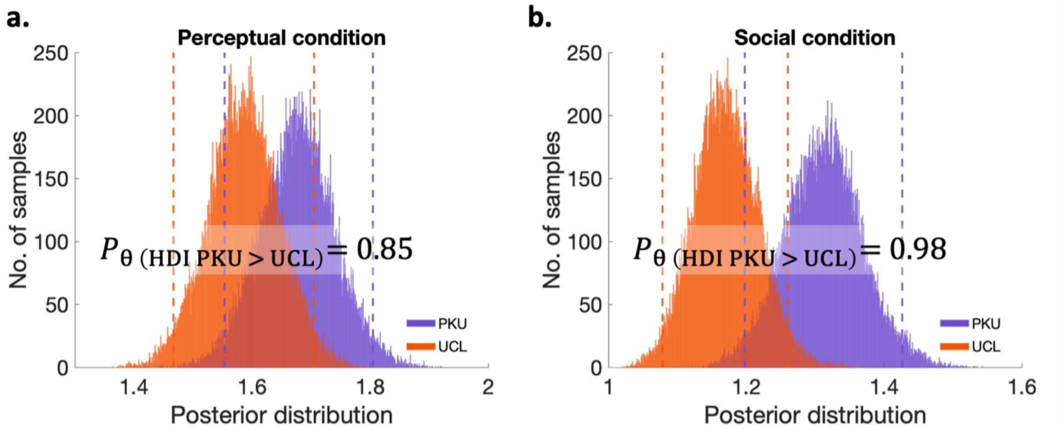


Figure 5.5: ***Metacognitive efficiency between sites in the perceptual (a) and social (b) conditions of Experiment 2.*** *Posterior distributions over group-level metacognitive efficiency for UCL and PKU participants separately. The dashed lines represent the 95% highest density intervals (HDI); $P\theta$ indicates the probability that the posterior samples of the PKU group are higher than the posterior samples of the UCL group.*

evidence on confidence and correlated this measure with main composite self-construal scores (SCS; [Singelis, 1994], IQ [Cattell, 1943] and cognitive insight (BCIS; [Beck et al., 2004]). Post-decisional processing did not correlate with main composite independent construal (perceptual condition: Pearson's r = 0.03, *P* = 0.73; social condition: Pearson's r = 0.09, *P* = 0.35). In line with recent findings that metacognitive sensitivity and IQ are independent constructs [Rouault et al., 2018], I also do not find any association between post-decisional processing and IQ (perceptual condition: Pearson's r = 0.09, *P* = 0.39; social condition: Pearson's r = 0.10, *P* = 0.32). The interaction between post-decision evidence and accuracy did not correlate with composite BCIS scores (perceptual condition: Pearson's r = 0.07, *P* = 0.47; social condition: Pearson's r = -0.08, *P* = 0.43). However, there was a positive correlation between post-decisional processing and both the BCIS self-certainty subscale (perceptual condition: Pearson's r = 0.25, *P* = 0.009; social condition: Pearson's r = 0.34, *P* = 4.25e-04) and the self-reflectiveness subscale (perceptual condition: Pearson's r = 0.23, *P* = 0.02; social condition: Pearson's r = 0.18, *P* = 0.06). When both sites were analyzed separately (N = 35 participants per site) there was a positive correlation between post-decisional processing and the self-confidence subscale in the PKU sample (PKU perceptual: Pearson's r = 0.25, *P* = 0.07, social: Pearson's r = 0.31, *P* = 0.02) but not in the UCL sample (UCL perceptual: Pearson's r = -0.03, *P* = 0.84, social: Pearson's r = 0.18, *P* = 0.20). Perhaps because of a lack of power to investigate individual differences in a sample of N = 53 participants per site, post-decisional processing was not correlated with the self-reflectiveness subscale in the samples taken individually (PKU perceptual: Pearson's r = 0.05, *P* = 0.71, social: Pearson's r = -0.17, *P* = 0.23); UCL perceptual: Pearson's r = 0.04, *P* = 0.79, social: Pearson's r = 0.06, *P* = 0.67).

5.4 Discussion

I here replicated and extended the results of the previous chapter by showing an enhanced susceptibility to post-decision evidence in participants with Chinese compared to British backgrounds. In the previous chapter I evidenced a metacognitive benefit led Chinese participants to become selectively more susceptible to new evidence following errors—a skill that led them to recognize and change their minds more often about errors. In this study I replicated this finding and also show the domain generality of this effect by leveraging a social form of new evidence, social advice. The cultural benefit of Chinese compared with British participants was commensurate for the social and non-social form of evidence alike.

The differences between cultural milieus in susceptibility to new evidence reported here complement and extend previous findings that Chinese populations are more affected by social influence than German and British populations [Korn et al., 2014, Mesoudi et al., 2015]. In particular, I suggest that such differences in a susceptibility to new evidence may be caused by heightened metacognition, rather than normative social compliance. In other words, recognizing the potential for error may prompt a search for corrective information from our peers. Notably, while Chinese participants were more susceptible to both social and perceptual forms of post-decision evidence, such effects were most prominent on trials where mistakes had been made. This interaction between the impact of post-decision evidence on confidence and accuracy is a key signature of metacognition [Fleming et al., 2018], and accordingly, the Chinese participants had consistently heightened metacognitive efficiency than UK participants in all three datasets. Our finding that cultural differences consistently, and selectively, occurred on error trials indicates that these cultural differences are primarily driven by metacognition, rather than a greater susceptibility to social influence irrespective of self-performance.

As perceptual post-decision evidence always disconfirmed a previous decision after errors (i.e., was always helpful), an alternative explanation of these findings is that PKU students simply processed disconfirming evidence to a greater extent than UCL students—in other words, they were less prone to confirmation bias [Talluri et al., 2018, Kappes et al., 2020]. However, additional analyses of the social task data nuance this interpretation. The social task allowed me to distinguish between cases of disagreement when advice was correct ('good advice') as well as when advice was wrong ('bad advice'). Notably, both PKU and UCL students were equally susceptible to bad advice that agreed with their wrong decision (suggesting similar susceptibility to confirmatory social information) and to bad advice that disagreed with their correct decision (suggesting similar susceptibility to social disagreement). Instead, differences between cultural backgrounds selectively manifested in a heightened susceptibility of PKU students to 'good' advice, even when it disagreed with their decision (**Figure 5.4**). This finding suggests that PKU students had heightened metacognitive evaluation of their performance, allowing them to selectively follow the advice when it is most beneficial.

Another line of evidence supporting a metacognitive explanation of these findings between sites is an association between an index of metacognitive processing (the tendency to specifically process new evidence on error trials) and an independent measure of cognitive insight (BCIS, [Beck et al., 2004]). PKU students had substantially higher baseline levels of self-reported cognitive insight than UCL students (**Supplementary Material 1.1**). In addition, inter-individual differences in cognitive insight, but not differences in sociocultural flexibility (as measured with the self-construal scale; [Choi et al., 2003]), predicted the degree of metacognitive processing in the sample as a whole (**Figure 5.1**).

I was also able to evaluate the domain-general nature of the cultural difference. On half of the trials post-decision evidence was perceptual, whereas on the other half it was presented as social advice. Differences between sites in post-decisional processing were similar across the social and perceptual forms of post-decision evidence, and the impact of both types of evidence was correlated across participants. Indeed, one interesting prediction of the cultural origins hypothesis of metacognition is that any cultural difference should be relatively domain-general, because the skills that are being acquired are metacognitive in nature rather than how to handle a particular type of information. A useful analogy is with the cultural acquisition of reading: even though a person might learn to read via information provided by others, they can subsequently apply that skill to read a variety of different books about topics that no longer have relevance for the social group. In this light, the finding that the impact of cultural variation on metacognitive efficiency generalizes to different types of evidence is expected from the theory.

Despite this similarity, participants at both sites adjusted their confidence levels to a lesser degree in response to social compared to perceptual evidence (**Figure 5.4a**), a difference that may have been due to the model generating simulated advisers with generally lower confidence levels than the participants. Whether social and perceptual evidence have a similar impact on post-decision processing when advisers' confidence is matched to that of the participant could be investigated in future experiments. Confidence is known to be closely linked to, and potentially informed by, variation in response times to make an initial decision [Rahnev and Fleming, 2019]. This finding suggests that the mechanism through which confidence is inferred may be different across cultures, but more research in this domain is needed to confirm this. Future studies could also seek to replicate these results using a confidence task without post-decision evidence, which I believe would give similar results [Rollwage et al., 2018]. Another limitation of this study is that neither the study described in chapter four nor five was pre-registered. Future studies should replicate the current findings in a larger sample and following pre-registration of hypothesized cultural differences.

In summary, across two behavioral experiments I demonstrate that East Asian participants (from Peking University) have heightened metacognitive evaluation of their task performance in comparison with Western participants (from University College London and New York University). These differences manifested in boosts to post-decisional processing following error trials, in the absence of differences in first-order performance. This pattern was also obtained in a new task where post-decision evidence was replaced with equivalent social advice, suggesting that culture shapes a domain-general tendency to evaluate and reflect on previous decisions. This provides a final piece of evidence that metacognition and mentalizing—in the form of being selectively more susceptible to more reliable advice—are shaped by our socio-cultural environment, a topic to which I will return in the General Discussion.

Chapter 6

General Discussion

6.1 Overview

Explicit metacognition, the ability to consciously reflect and report on other cognitive processes, is considered a uniquely human faculty [Frith, 2012]. Homo sapiens' widespread knowledge and skill is largely owing to our ability to translate internal insights into words that others can understand [Frith, 2012, Shea et al., 2014]. While the evolutionary reason for the development of metacognition may have been primarily social, we know little about what role social information plays in metacognitive processes.

Across a number of studies, I show that metacognitive efficiency is shaped by our social environment. One prominent theory in my research has been the mentalizing-is-prior theory, which suggests that metacognition is realized by turning our understanding of other people's mental states to ourselves. This hypothesis posits that people do not have privileged and direct access to their own minds, but instead, infer their own thoughts and feelings indirectly, as they would infer the mental states of others.

In chapter two, I tested this theory in the realm of mentalizing, the set of cognitive processes involved in inferring other people's mental states, and measured both the metacognitive and mentalizing efficiency of a general population dataset [N=477]. In line with my pre-registered hypotheses, I found that mentalizing and metacognitive efficiency were positively correlated, even after controlling for first order performance, IQ, age, gender and education. By modelling the trial-by-trial formation of confidence I showed that mentalizing efficiency predicted the association between response times and confidence, suggesting those with better mentalizing efficiency were more sensitive to inferential cues to self-performance. Because response time is an indirect cue for confidence that is similarly used to infer the confidence of someone else [Patel et al., 2012], this finding suggests that mentalizing facilitates metacognition by allowing people to apply the same strategies to read their own minds as they would to read those of others.

The mentalizing-is-prior theory also reinterprets Autism Spectrum Condition (ASC) as a meta-representational disturbance [Baron-Cohen et al., 1985, Frith and Happé, 1999]. Specifically, it predicts that the mentalizing difficulties that characterize this condition should also hinder the development of metacognition. In chapter two I indeed found that a group of autistic participants had lower metacognitive efficiency and mentalizing efficiency than age, gender, education and IQ matched controls. One particular aspect of metacognition that was compromised in ASC versus a comparison population was the use of response times in recognizing committed errors. Given that we had previously shown this propensity to be related to mentalizing efficiency, these results suggest that autistic difficulties with understanding other people's mental states may also have downstream consequences on the capacity to apply that same understanding to oneself.

In the remainder of the thesis, I built upon this idea to test if metacognitive efficiency is sensitive to cultural norms that prioritize the group over the individual. Metacognition was quantified as an adaptive boost to processing new evidence following errors in participants with matched age, gender, IQ and family income but distinct cultural backgrounds (British/American or Chinese). In chapter three I found that Chinese participants have heightened metacognitive evaluation of their task performance in comparison with British and North American participants. This effect was driven by Chinese participants selectively processing new evidence to correct an ongoing error more than their Western counterparts. In chapter four I replicated this effect in a new dataset and show that Chinese participants have both enhanced metacognitive and mentalizing efficiency in the way in which they process social advice in comparison with British participants. Strikingly, this cultural difference was found irrespective of whether the new evidence was provided in a perceptual or social format, suggesting that the information benefit (and not social pressure to comply) was driving the reported cross-cultural differences.

Together, these results suggests that one common meta-representational mechanism, which is similarly involved in reading the minds of oneself and others, is sensitive to social communication about internal states and collaborative environments. Here, I will review these findings in light of previous research and discuss its implications.

6.2 Social shaping of mentalizing and metacognition

In this thesis I tested the proposal that the capacity to understand other people's thoughts facilitates the recognition of those same thoughts in oneself. One laboratory study, which quantified metacognitive efficiency as the trial-by-trial association between confidence and accuracy [Fleming et al., 2010, Fleming and Lau, 2014, Frith and Frith, 2012, Yeung and Summerfield, 2012], has shown, in line with this view, that people with better metacognitive efficiency also tend to score better on mentalizing tasks [Nicholson et al., 2020] (see: [Carpenter, 2000, Nicholson et al., 2019] for different results). An implication of this finding is that having restricted access to the mental states of others should hinder the development of metacognitive efficiency too. Some studies have tested this implication in the realm of ASC and indeed found commensurate metacognitive and mentalizing problems in ASC [Grainger et al., 2016a, Nicholson et al., 2020, Williams et al., 2018]. However, other studies have found inconsistent results

[Carpenter et al., 2019a, Wojcik et al., 2013], suggesting that it is, so far, unknown whether the development of metacognition is dependent on exposure to social information about other people's mental states.

One difficulty with interpreting prior work is that the measurement of metacognitive processes is often confounded by the first-order thinking that it evaluates, which itself may vary across individuals and clinical groups. For example, the Goodman-Kruskall gamma coefficient between trial-by-trial accuracy and confidence [Nelson, 1984] is confounded by *type-1 sensitivity* and *metacognitive bias* [Fleming and Lau, 2014, Maniscalco and Lau, 2012, Maniscalco and Lau, 2014, Masson and Rotello, 2009, Rahnev and Fleming, 2019]. The impact of this confound may be particularly pertinent to studies comparing autistic and comparison populations, as sensory (over-) sensitivity [Ewbank et al., 2016, Lieder et al., 2019, Pirrone et al., 2017] and over-confidence [McMahon et al., 2016, Milne et al., 2002, Zalla et al., 2015] are sometimes found to be higher in autistic than in comparison groups. In other words, previously reported measures of metacognitive sensitivity may have been confounded by altered sensory sensitivity in autistic participants.

Another caveat is that previous studies have focused on an association between average metacognitive efficiency and mentalizing efficiency across trials. Much less is known about how mentalizing efficiency affects the computational processes that give rise to trial-by-trial fluctuations in confidence. Work in cognitive psychology has often shown that people have poor access to the reasons for their actions but instead infer these from contextual cues (even if these cues are experimentally decoupled from the true underlying intention; [Gazzaniga, 1995, Gazzaniga, 2000, Wegner, 2002, Wilson, 2002]). For example, when asked to rate their confidence in a previous decision, people's confidence reports may be affected by various (behavioural) cues that are more or less related to the decision, such as response times [Kiani et al., 2014, Patel et al., 2012], social context [Bahrami et al., 2010, Bang et al., 2017, van der Plas et al., 2022], as well as the quantity and reliability of evidence [Campbell-Meiklejohn et al., 2010, De Martino et al., 2017, Kiani and Shadlen, 2009, Pleskac and Busemeyer, 2010, Campbell-Meiklejohn et al., 2010]. Intriguingly, response times have a causal impact on the confidence levels people ascribe not only to themselves, but also to others [Palser et al., 2018, Patel et al., 2012], suggesting common cues may be recruited in both cases. If the Mentalizing-is-prior view is correct, mentalizing efficiency should facilitate the indirect inference of confidence people make from their behavioural cues.

In chapter two I provided empirical support for the mentalizing-is-prior theory. Using HMeta-d hierarchical regression models, I found that participants with better mentalizing efficiency also had better metacognitive efficiency, even after controlling for sensory sensitivity and metacognitive bias. My findings go beyond estimating correlations between metacognition and mentalizing, and reveal a potential route through which mentalizing may affect metacognitive processes. Specifically, a greater ability to identify the mental states of others from behavioral cues may also allow people to identify those same cues within themselves. This finding indicates, in line with the mentalizing-is-prior view, that metacognition and mentalizing may have a common computational basis—allowing people to focus more on cues that are predictive of one's own and other people's mental states.

One intriguing outcome was that I did not find poorer metacognitive efficiency in people with higher scores of autistic traits, even though such a correlation has been found elsewhere [Nicholson et al., 2020]. An explanation for this discrepancy is that, on their metacognition task, Nicholson and colleagues rewarded high confidence on correct trials and low confidence on error trials. In other words, better metacognition was explicitly incentivized. This may have created a disadvantage for autistic participants, who have to difficulties interpreting and learning from (ambiguous) feedback [Broadbent and Stokes, 2013, Greene et al., 2019, Reed, 2019, Robic et al., 2015, Sapey-Triomphe et al., 2018, Zwart et al., 2018]. A second explanation is that variation in autistic traits in the general population may not have been pronounced enough to find statistically significant differences in metacognitive efficiency, as hinted by the more pronounced differences in metacognitive efficiency found in chapter three.

Notably, in an exploratory set of analyses I modeled the shared co-variance between mentalizing and metacognitive efficiency using structural equation modelling. The rationale for this approach is that metacognitive and mentalizing efficiency are both noisy measurements of which only their shared co-variance may pertain to a true meta-representational faculty. In support of this hypothesis, I found that the association between autistic traits and a shared, meta-representational faculty was stronger than that between autistic traits and metacognition and mentalizing efficiency separately (**Figure 2.6**).

Another notable finding was that those with greater self-reported difficulties with social communication and social understanding also had poorer metacognitive efficiency. This finding relates to the mentalizing-is-prior prediction that having restricted access to social communication and difficulties with understanding other people's thoughts and actions have downstream consequences for the development of metacognition. But is this effect causal, i.e. does being faced with social interaction strengthen metacognitive efficiency?

Some initial evidence suggests this is the case. For instance, a key predictor of mentalizing efficiency in children is the amount of social communication a child is exposed to [Dunn et al., 1991, Dunn and Brown, 1993, Jenkins and Astington, 1996, Brown et al., 1996]. Further, developmental work has shown that societal norms and conditions that obstruct social communication and explanations of other people's mental states affect the development of mentalizing in children [Liu et al., 2008, Peterson and Siegal, 1995, Richardson et al., 2020, Wellman and Liu, 2004]. On the basis of these findings, one could predict that people who are more exposed to collaborative environments also have better metacognitive efficiency. In chapter three and four I compared the metacognitive profiles of adults that were raised in highly collaborative environments with those that were raised in less collaborative environments and found support for this hypothesis.

In summary, these results suggest that the capacity to understand our own mental states is similar to that involved in understanding the mental states of others. My results suggest that mentalizing efficiency impacts the association between confidence and response times—a behavioural cue that is similarly used to infer confidence of self and others. Together, these findings suggest that metacognition is at least partially an indirect process, “read out” from behavioural cues. The personal examination of one’s own conscious thoughts and processes is a central topic of philosophical debate, as prominently exemplified by Descartes’ famously relying on introspection to deduct his own existence (“Cogito, ergo sum” [Descartes, 1644]). This work suggests that research on the social basis of metacognition forms a useful step towards the modern investigation of this classic phenomenon.

6.3 Autism as a metacognitive condition

Chapter three posed the question whether people with mentalizing difficulties are similarly restricted in their metacognitive capacities, which I tested in the realm of ASC. This study built directly upon the insignificant correlation between autistic traits and metacognitive efficiency found in chapter two. I reasoned that, if this null finding was driven by autistic traits in the general population not being pronounced enough to find statistically significant differences in metacognitive efficiency, metacognitive differences may be more visible in a clinically diagnosed group of autistic individuals. The method employed in chapter three was similar to that of chapter two.

In line with my pre-registered hypotheses, I found that metacognitive efficiency was significantly lower in the autism group than in the comparison group, in the absence of any differences in IQ, age, gender and education. An interesting feature of this finding is that the effect size of the frequentist comparison was larger than that of the Bayesian analysis. One explanation for this may be that metacognitive efficiency in autism may not be weaker on average, but rather more polarized (extremely low or extremely weak) than that of comparison participants

[Pariser, 1981, Shields-Wolfe and Gallagher, 1992], which is a type of variability to which Bayesian analyses are more sensitive. More research is needed to better understand whether autistic participants indeed have more polarized differences in metacognition, and if so, whether this is attributed to autistic people engaging in alternative, maybe even more cognitively demanding, processes to compensate for their metacognitive difficulties [Livingston et al., 2019a, Livingston et al., 2019b].

How do these findings relate to what we know about ASC? Autistic lives can be very different from the lives of typical adults. Estimates of autistic people who can live autonomously with usual levels of support range from 4% [Howlin et al., 2004] to 64% [Cederlund et al., 2008]. Employment at either full or part-time basis is similarly low in autistic adults, ranging from around 7% to 40% [Engström et al., 2003, Helles et al., 2017]. Moreover, 17% of autistic adults live independently and 53% has had paid employment, these percentages are lower for people on the autism spectrum than for adults with learning disability (66% for independent living, and 88% for paid employment respectively) or intellectual disabilities (34% for independent living, and 63% for paid employment, respectively; [Anderson et al., 2014]). However, within the autistic population that participates in laboratory studies, studies of perception [Greimel et al., 2013, Peiker et al., 2015, Plaisted et al., , Powell et al., 2016] and learning [Brown et al., 2010, D'Cruz et al., 2016, Luman et al., 2009] have rarely found differences between autistic and comparison participants. It is important to note that the participant samples of these studies is not necessarily representative of the autistic population. First, the tested samples were often largely–or exclusively–male. This is problematic as recent estimates of the male to female ratio in autism is around 3:1 [Loomes et al., 2017]. Second, the vast majority of studies set a minimum IQ, whereas a significant proportion of autistic people also have a co-occurring intellectual disability [estimated to be between 50-55%, Charman et al., 2011]. This suggests that the results of some empirical studies may not generalise to the wider autistic population and/or that decision-making in real life is intrinsically different from laboratory decision-making tasks. One possibility is that general cognition is typical in some autistic individuals, but that the way in which cognition itself is (hierarchically) evaluated may be atypical [Friston et al., 2013, Palmer et al., 2017]. In everyday situations, where decisions are often subjective, feedback and clear rules on how to decide are often lacking. Whilst on these laboratory decision-making tasks or structured assessments, an objectively 'correct' answer can often be learned over the course of multiple decisions. This idea fits with theoretical proposals that especially subjective and metacognitive aspects of cognition are compromised in ASC [Carruthers, 2009, Frith and Happé, 1999, Happé and Frith, 2006].

Even if metacognitive processes are largely independent from first-order cognition, such as perceptual ability [Fleming et al., 2010, Maniscalco and Lau, 2012, Maniscalco and Lau, 2014], the ability to learn associations [Rouault et al., 2019] and fluid intelligence [Rouault et al., 2018]; metacognition plays a role in several types of decisions that play an important role in everyday scenarios, such as value based decision-making (deciding on the basis of subjective preferences; [De Martino et al., 2013, Woodcock et al., 2020]), decisions that are based on emotions and bodily arousal [Bird and Cook, 2013, Kuzmanovic et al., 2019, Shah et al., 2016], and decisions made in social contexts [Bahrami et al., 2010, Bahrami et al., 2012, Bang et al., 2017]. In addition, some work suggests that metacognition is more predictive of everyday indices of successful decision-making—such as educational [Bakracevic Vukman and Licardo, 2010, Isaacson and Fujita, 2006, Narang and Saini, 2013] and vocational success [Cho and Linderman, 2019, Marshall-Mies et al., 2000]—than reward learning and perceptual decision-making paradigms alone [Koren et al., 2006]. For example, under-confidence in one's performance can cause shifts in perceptual sensitivity; and if people are insensitive to their own feelings, this could cause them to appear apathetic or repetitive on the outside (interestingly both traits that are commonly attributed to ASC [Howlin et al., 2013, Robinson et al., 2016, Warrier et al., 2019]). My results suggest that, not only may some cognitive difficulties in ASC be explained by a metacognitive deficit, also seemingly "non-social" spectrum phenotypes may be a by-product of having less access to social information (in addition to already existing genetic predispositions: [Heyes and Frith, 2014]). Naturally, more work in this domain is needed to evaluate how the development of metacognition and mentalizing interact in ASC.

It is important to note the limitations of this study, of which the most important one is its representativeness. Our autistic sample was largely female whereas estimates of the male to female ratio in autism is around 3:1 [Loomes et al., 2017]. Second, our autistic participants all had fluid intelligence scores similar to those of the general population dataset, whereas a significant proportion of autistic people also have a co-occurring intellectual disability (estimated to be between 50-55% [Charman et al., 2011, Loomes et al., 2017]). In a recent study I employed a literature review approach to assess if metacognitive processes are compromised in ASC. By reviewing 74 empirical studies [N=5,111 total participants, N=1,932 autistic and N=3,179 comparison] published between 1998 to 2020, I found that autistic participants have compromised performance on decisions that rely on metacognitive processes (such as decisions that require evaluating one's own or other people's knowledge or preferences). On the other hand, autistic participants perform similarly to non-autistic participants on tasks that rely primarily on first-order processes (such as perceptual decision making or reward learning; [Van der Plas, Mason et al., 2022]). Together with the findings presented in chapter three, these findings suggest that metacognitive difficulties in autism may generalize to a broader set of decision making problems that are common in ASC.

One potential application of this work is to provide more targeted support. If the everyday problems autistic people face mostly arise at a metacognitive level of cognition, this may prompt caregivers to provide support at especially this level of cognition. Recent work is testing whether metacognitive training paradigms are found to be helpful and to whom [Maras et al., 2017, Lamash and Josman, 2021]. In schizophrenia, the positive impact of metacognitive training paradigms generalize to mentalizing efficiency [Lysaker et al., 2012, Lana et al., 2017], supporting my findings that mentalizing and metacognition have a common cognitive basis. One striking feature of these metacognitive trainings is that they do not give people repeated feedback on their confidence estimates (as has been done in other trainings [Carpenter et al., 2019a]). Instead, these trainings facilitate group discussions of mental states. These results suggests that the mere act of talking about feelings may positively impact metacognitive efficiency. More research on whether the mere act of talking about feelings can causally benefit metacognitive efficiency in adults is needed. One route to explore this question is by comparing groups of populations that have different norms of social communication and collaboration, as was done in the next chapter of my thesis.

6.4 Cultural contributions to metacognition

The previous chapter evaluated how developmental disturbances that give rise to compromised mentalizing efficiency may also affect metacognitive efficiency. This finding, together with the finding from chapter two that individual trait differences in problems with social communication and understanding negatively correlate with metacognitive efficiency, suggest that metacognition is realized by turning an understanding of other people's mental states to oneself. In other words, when a developmental condition prevents autistic people from understanding the feelings of others, this can impact their metacognitive development too. Beyond looking at these types of *internal* obstructions to social communication and explanations of other people's mental states (e.g., a clinical inability to grasp the mental states of others), there exist also *external* obstructions to social communication and explanations of other people's mental states (e.g., living in a society where talking about mental states is considered impolite). A number of studies have explored these external obstructions and suggest that cultural obstructions to accessing information about other people's mental states negatively impact and/or delay the development of mentalizing [Liu et al., 2008, Peterson and Siegal, 1995, Richardson et al., 2020, Wellman and Liu, 2004].

This cultural origins hypothesis posits that metacognitive efficiency is susceptible to differences in our socio-cultural environment [Heyes et al., 2020, Heyes, 2018, Proust and Fortier, 2018]. There are several reasons why comparing Chinese and British or American cultures may provide a promising testbed to investigate this theory empirically. First of all, people that were raised in China are often described as being more collectivist in that they traditionally emphasize harmony with others, whereas people who were raised in Northern Europe or the United States are commonly viewed as more individualistic [Hofstede, 2011]. Moreover, Chinese populations are more likely to pay attention and conform to others' opinions than UK or US populations [Korn et al., 2014, Mesoudi et al., 2015, Oeberst and Wu, 2015]. This is relevant, as advice taking is in many ways similar to metacognition, as both processes involve processing new evidence after an initial decision has already been made ([van der Plas et al., 2019]). Finally, Chinese populations are thought to be more inter-dependent in their thinking styles than Westerners [Singelis, 1994]. In light of the mentalizing-is-prior theory, a greater tendency for collaboration and having shared goals could give rise to metacognitive differences too, but whether this is actually the case was hitherto unknown.

Some cross-cultural studies have shown that confidence bias differs between Chinese and UK or US populations. A general finding in this literature is that Chinese participants report higher confidence in their own responses than UK or US populations [Moore et al., 2018, Yates et al., 1989, Yates et al., 1998]. Two studies have extended this work to investigate whether metacognitive efficiency is similarly different between Chinese and UK or US population, but found inconsistent results [Yates et al., 1989, Yates et al., 1998]. One potential issue with this work is that the researchers left first-order performance (choice accuracy) free to vary across individuals, which may have confounded the results as people tend to be better at discriminating their own performance when they perform better on a task [Fleming and Lau, 2014]. As such, it is currently unknown whether metacognitive efficiency also varies between cultures when the metric of metacognitive efficiency is not confounded by differences in choice accuracy.

In addition, it is possible that not metacognition per se, but the computations involved in the construction of confidence, vary across cultures. It could be, for example, that different educational settings impact distinct ways of mental state evaluation, leading to differences in the process (rather than the mere outcome) of metacognition [Liu et al., 2008]. For instance, confidence ratings are informed by evidence that becomes available after deciding ("post-decision evidence"), where it allows the decider to rate higher confidence when the new evidence is confirming, and lower confidence when the new evidence is disconfirming [Fleming et al., 2018]. It could be that different cultures specifically modulate susceptibility to post-decision evidence but not metacognitive efficiency per se.

In chapter three, I directly tested the Cultural Origins Hypothesis. I recruited two groups of participants that were matched in terms of age, gender, IQ and family income but had different cultural backgrounds. Chinese nationals who had not lived abroad for longer than six months were recruited from Peking University (PKU, Beijing, China); British nationals who had not lived abroad for longer than six months were recruited from University College London (UCL, London, UK). To ensure first-order performance was matched, I selected three levels of evidence strength that were of subjectively similar perceptual strength across participants. As a result of this calibration, first order task difficulty was matched across individuals and sites. In line with prior work, I hypothesised that PKU participants would exhibit a greater sensitivity to new evidence than UCL participants. Strikingly, I found this enhanced susceptibility to new evidence of PKU participants only on error-trials, suggesting enhanced metacognitive efficiency rather than a greater susceptibility to new evidence in general.

In exploratory analyses I investigated if there is a specific personality trait that predicts an enhanced sensitivity to new evidence following errors. If cultural norms of harmony and collaboration strengthen metacognition, one would predict that specifically those with a greater desire for social harmony and inter-dependency would exhibit a sensitivity to new, correcting, evidence. Importantly however, I did not find this. The only consistently different personality trait between Chinese and British participants was a difference in cognitive insight, a self-reported ability to differentiate reality from subjective experience [Beck et al., 2004]. In turn, individuals who scored higher on cognitive insight were also more susceptible to new evidence on error trials. This suggests that the Chinese culture may exert its influence on metacognitive processing directly, without the interference of cultural norms of collaboration.

6.5 Advice taking as a model of metacognition and mentalizing

In chapter three I showed that people who were raised in cultures that emphasize the collective and social harmony over the individual are more likely to selectively process new perceptual evidence after having made an error than people who were raised in more individualistic cultures. This finding suggests that the ability for metacognition is susceptible to cultural differences and is in line with the Cultural Origins Hypothesis [Heyes et al., 2020]. This hypothesis suggests that, just like how the skill to read can be applied to different books, the ability for mindreading can be similarly applied to read one's own and others' minds [Heyes, 2018], which was a key prediction for chapter four.

In chapter three I studied whether cultural background affects how new *perceptual* evidence is processed. In chapter four I build upon this by testing whether cultural differences in post-decision evidence processing is indifferent to the type of evidence at hand (i.e., *social or perceptual*). This manipulation of evidence type also allowed me to explore a second prediction of the mentalizing-is-prior theory. Namely, that the process through which people infer the mental states of others is similar to the process employed to infer one's own mental states.

To test these predictions, I recruited two new samples of participants from both PKU and UCL, again ensuring that age, gender, IQ and family income were matched across groups. Participants made perceptual decisions based on pre-decision evidence of varying strength (weak, medium or strong), similar to the perceptual task. However, on half of the trials, perceptual post-decision evidence was replaced with the confidence and perceptual judgment of another agent. This advice was generated from a generative model that had the same perceptual sensitivity as the participant, and therefore, had the same accuracy and confidence level as the participant. This allowed me to investigate the impact of social and perceptual post-decisional evidence on confidence ratings, and compare their impact across cultures.

When I compared the propensity to inform confidence on the basis of new advice, I found that the confidence ratings of PKU participants were more affected not only by perceptual but also social post-decision evidence. Similar to perceptual post-decision evidence, this enhanced susceptibility to social advice in PKU versus UCL participants was restricted to trials on which the advice was correct, and the participant wrong. On one hand, this finding is a conceptual replication of the findings from chapter tree, indicating that PKU participants had better metacognitive efficiency in knowing when advice was most beneficial (i.e., when they were initially wrong). On the other hand, this finding shows that PKU participants were more susceptible to social advice only when the advice was correct (i.e., and not when both themselves and the adviser were wrong). This suggests that beyond a metacognitive benefit, PKU participants also had better *mentalizing efficiency*, i.e., a greater ability to differentiate between the advisers' correct and erroneous advice with advice susceptibility. This final piece of evidence suggests that the mentalizing-is-prior hypothesis is the most parsimonious explanation of the relationship between metacognitive and mentalizing efficiency.

In conclusion, advice-taking is a common form of knowledge sharing that has widespread implications, from clinical problems with treatment compliance to the way academics synthesize and share their acquired knowledge. People often update their beliefs based on others' opinions, such as experts, friends, family, and online users [Bonaccio and Dalal, 2006]. This work suggests that advice-taking is a complex problem that may rely both on metacognitive efficiency and mentalizing efficiency. A challenge for future work remains identifying which aspects facilitate good knowledge sharing, to ensure knowledge sharing in humans is as effective, and fair, as possible.

6.6 General Conclusion

Across four experiments on nearly a thousand participants, I evidenced that metacognition—the ability for "thinking about thinking", such as self-doubt or hesitation, is deeply rooted in our social environment. In chapter two, I showed that similar neurocomputational processes are involved in understanding one's own and other people's minds (mentalizing efficiency). Strikingly, both inferences of mind were made, at least in part, indirectly—by relying on cues rather than direct introspection. This suggests that access to other people's explanations of their mental states shapes an ability to recognize similar thought processes in oneself. This idea was further tested in chapter three, where I found that people who have restricted access to other people's minds (as is the case in Autism Spectrum Condition) also tended to have difficulties with metacognition. These studies suggest that subtle differences in socio-cultural context can affect the method and efficiency with which people perceive their own minds. This was confirmed in chapter four, where I compared groups of people with different cultural backgrounds. I showed that in cultures where collaboration is emphasized, metacognitive efficiency is better than in cultures where working alone is the norm. In chapter five, I built upon this finding by showing that socio-cultural differences have a domain-general impact on metacognition—affecting the integration of social and non-social types of evidence to a similar extent.

These results support the mentalizing-is-prior theory. Compared with alternative theories on how the human meta-representational system may be organised (Introduction), this theory predicts, together with the mentalizing- and metacognition-is-prior theory, that the metacognitive and mentalizing reasoning process is similar—a prediction for which I found support in chapters two, three and five. In addition, the finding that metacognitive efficiency is shaped or affected by differences in social communication (for which I found evidence in chapter two, three, four and five), nuances this further by undermining the metacognition-is-prior theory—under which direct inference would make metacognitive efficiency insensitive to external differences—more than the mentalizing-is-prior theory. However, because all the data collected in my studies was correlational in nature, the causal predictions that the mentalizing-is-prior theory makes remains an avenue for future research. In addition, these theories are all models (and are also tested as such in chapter two) and are, therefore, inherently wrong. The actual meta-representational reasoning process may be a more nuanced version of either view. For example, it could be that a hybrid meta-representational system starts as a two-mechanisms system but later develops into either a mentalizing- or metacognition-is-prior system. The way in which it develops may depend on whether one's socio-cultural environment either emphasizes one's own or other people's thoughts—"pruning" the system towards either a metacognition- or mentalizing-is-prior system. Future studies could collect longitudinal data to reveal how the development of metacognition and mentalizing interact.

In practice, these results highlight the importance of understanding our own and other minds for effective and fair information sharing. Some of the most prominent societal issues arise from difficulties at this level. If advice-taking is the cumulative process of sharing unique worldviews with others—what happens during the translational turn from thought to word is key. Not knowing the reliability of one's own thought processes and over-confidence are likely to play a role in the distribution of misinformation; not knowing the reliability of other people's thought processes may be the root of epistemic injustice. Perhaps a useful next step is to seek solutions to these problems in strengthening what led to the development of the human capacity for metacognition in the first place: our social environment. On the basis of the findings presented in this thesis, I expect that this future work will confirm what my thesis suggests:

In multis versor, ergo sum

—I socialize, therefore I am

Bibliography

[Abell et al., 2000] Abell, F., Happé, F., and Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, 15(1):1–16.

[Achával et al., 2010] Achával, D. d., Costanzo, E. Y., Villarreal, M., Jáuregui, I. O., Chiodi, A., Castro, M. N., Fahrer, R. D., Leiguarda, R. C., Chu, E. M., and Guinjoan, S. M. (2010). Emotion processing and theory of mind in schizophrenia patients and their unaffected first-degree relatives. *Neuropsychologia*, 48(5):1209–1215.

[Adolphs, 2009] Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual review of psychology*, 60:693–716.

[Allison et al., 2012] Allison, C., Auyeung, B., and Baron-Cohen, S. (2012). Toward Brief "Red Flags" for Autism Screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 Cases and 3,000 Controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(2):202–212.e7.

[Anderson et al., 2014] Anderson, K. A., Shattuck, P. T., Cooper, B. P., Roux, A. M., and Wagner, M. (2014). Prevalence and correlates of postsecondary residential status among young adults with an autism spectrum disorder. *Autism*, 18(5):562–570.

[Andrew Gelman and Donald B. Rubin, 1992] Andrew Gelman and Donald B. Rubin (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

[Ashwood et al., 2016] Ashwood, K. L., Gillan, N., Horder, J., Hayward, H., Woodhouse, E., McEwen, F. S., Findon, J., Eklund, H., Spain, D., Wilson, C. E., Cadman, T., Young, S., Stoencheva, V., Murphy, C. M., Robertson, D., Charman, T., Bolton, P., Glaser, K., Asherson, P., Simonoff, E., and Murphy, D. G. (2016). Predicting the diagnosis of autism in adults using the Autism-Spectrum Quotient (AQ) questionnaire. *Psychological medicine*, 46(12):2595–2604.

[Bahrami et al., 2012] Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., and Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1):3–8.

[Bahrami et al., 2010] Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., and Frith, C. D. (2010). Optimally interacting minds. *Science (New York, N.Y.)*, 329(5995):1081–1085.

[Bakracevic Vukman and Licardo, 2010] Bakracevic Vukman, K. and Licardo, M. (2010). How cognitive, metacognitive, motivational and emotional self-regulation influence school performance in adolescence and early adulthood. *Educational Studies*, 36(3):259–268. Publisher: Routledge.

[Balsdon et al., 2020] Balsdon, T., Wyart, V., and Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, 11(1):1753.

[Bang et al., 2017] Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y. F., Latham, P. E., Bahrami, B., and Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, 1(6):0117.

[Bang et al., 2014] Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y., Roepstorff, A., Rees, G., Frith, C. D., and Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition*, 26:13–23.

[Baron-Cohen, 1992] Baron-Cohen, S. (1992). Out of Sight or Out of Mind? Another Look at Deception in Autism. *Journal of Child Psychology and Psychiatry*, 33(7):1141–1155.

[Baron-Cohen et al., 1992] Baron-Cohen, S., Allen, J., and Gillberg, C. (1992). Can Autism be Detected at 18 Months?: The Needle, the Haystack, and the CHAT. *British Journal of Psychiatry*, 161(6):839–843. Publisher: Cambridge University Press.

[Baron-Cohen et al., 1985] Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a “theory of mind” ? *Cognition*, 21(1):37–46.

[Baron-Cohen et al., 1986] Baron-Cohen, S., Leslie, A. M., and Frith, U. (1986). Mechanical, behavioural and Intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology*, 4(2):113–125. Place: United Kingdom Publisher: British Psychological Society.

[Baron-Cohen et al., 2001] Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and 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, and allied disciplines*, 42(2):241–251.

[Bartsch and Wellman, 1995] Bartsch, K. and Wellman, H. M. (1995). *Children talk about the mind.* Children talk about the mind. Oxford University Press, New York, NY, US. Pages: viii, 234.

[Beck et al., 2004] Beck, A. T., Baruch, E., Balter, J. M., Steer, R. A., and Warman, D. M. (2004). A new instrument for measuring insight: the Beck Cognitive Insight Scale. *Schizophrenia Research*, 68(2):319–329.

[Becker et al., 2014] Becker, M. P. I., Nitsch, A. M., Schlösser, R., Koch, K., Schachtzabel, C., Wagner, G., Miltner, W. H. R., and Straube, T. (2014). Altered emotional and BOLD responses to negative, positive and ambiguous performance feedback in OCD. *Social Cognitive and Affective Neuroscience*, 9(8):1127–1133.

[Behrens et al., 2008] Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., and Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, 456(7219):245–249.

[Bertrams, 2021] Bertrams, A. (2021). Internal reliability, homogeneity, and factor structure of the ten-item Autism-Spectrum Quotient (AQ-10) with two additional response categories. *Experimental Results*, 2:e3.

[Bhaskar and Thomas, 2019] Bhaskar, V. and Thomas, C. (2019). The Culture of Overconfidence. *American Economic Review: Insights*, 1(1):95–110.

[Bird and Cook, 2013] Bird, G. and Cook, R. (2013). Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism. *Translational Psychiatry*, 3(7):e285–e285.

[Bollen and Pearl, 2013] Bollen, K. A. and Pearl, J. (2013). Eight Myths About Causality and Structural Equation Models. In Morgan, S. L., editor, *Handbook of Causal Analysis for Social Research*, pages 301–328. Springer Netherlands, Dordrecht.

[Bonaccio and Dalal, 2006] Bonaccio, S. and Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2):127–151.

[Brewer et al., 2017] Brewer, N., Young, R. L., and Barnett, E. (2017). Measuring theory of mind in adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 47(7):1927–1941.

[Britten et al., 1992] Britten, K. H., Shadlen, M. N., Newsome, W. T., and Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 12(12):4745–4765.

[Broadbent and Stokes, 2013] Broadbent, J. and Stokes, M. A. (2013). Removal of negative feedback enhances WCST performance for individuals with ASD. *Research in Autism Spectrum Disorders*, 7(6):785–792.

[Bronfman et al., 2015] Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., and Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society B: Biological Sciences*, 282(1810):20150228. _eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2015.0228.

[Brown et al., 2010] Brown, J., Aczel, B., Jiménez, L., Kaufman, S. B., and Grant, K. P. (2010). Intact implicit learning in autism spectrum conditions. *Quarterly Journal of Experimental Psychology*, 63(9):1789–1812.

[Brown et al., 1996] Brown, J. R., Donelan-McCall, N., and Dunn, J. (1996). Why talk about mental states? The significance of children's conversations with friends, siblings, and mothers. *Child Development*, 67(3):836–849. Place: United Kingdom Publisher: Blackwell Publishing.

[Burke et al., 2010] Burke, C. J., Tobler, P. N., Baddeley, M., and Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, 107(32):14431.

[Cahill et al., 2020] Cahill, V. A., Malouff, J. M., Little, C. W., and Schutte, N. S. (2020). Trait perspective taking and romantic relationship satisfaction: A meta-analysis. *Journal of Family Psychology*, 34(8):1025–1035. Place: US Publisher: American Psychological Association.

[Campbell-Meiklejohn et al., 2017] Campbell-Meiklejohn, D., Simonsen, A., Frith, C. D., and Daw, N. D. (2017). Independent Neural Computation of Value from Other People's Confidence. *The Journal of Neuroscience*, 37(3):673.

[Campbell-Meiklejohn et al., 2010] Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., and Frith, C. D. (2010). How the Opinion of Others Affects Our Valuation of Objects. *Current Biology*, 20(13):1165–1170.

[Carpenter et al., 2019a] Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., and Fleming, S. M. (2019a). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of experimental psychology. General*, 148(1):51–64.

[Carpenter et al., 2019b] Carpenter, K. L., Williams, D. M., and Nicholson, T. (2019b). Putting Your Money Where Your Mouth is: Examining Metacognition in ASD Using Post-decision Wagering. *Journal of Autism and Developmental Disorders*, 49(10):4268–4279.

[Carpenter, 2000] Carpenter, W. T. (2000). Decisional Capacity for Informed Consent in Schizophrenia Research. *Archives of General Psychiatry*, 57(6):533–538.

[Carruthers, 2009] Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, page 62.

[Cattell, 1943] Cattell, R. B. (1943). The description of personality: basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, 38(4):476–506.

[Cederlund et al., 2008] Cederlund, M., Hagberg, B., Billstedt, E., Gillberg, I. C., and Gillberg, C. (2008). Asperger Syndrome and Autism: A Comparative Longitudinal Follow-

Up Study More than 5 Years after Original Diagnosis. *Journal of Autism and Developmental Disorders*, 38(1):72–85.

[Chantiluke et al., 2015] Chantiluke, K., Barrett, N., Giampietro, V., Brammer, M., Simmons, A., Murphy, D. G., and Rubia, K. (2015). Inverse Effect of Fluoxetine on Medial Prefrontal Cortex Activation During Reward Reversal in ADHD and Autism. *Cerebral Cortex*, 25(7):1757–1770.

[Charles et al., 2014] Charles, L., King, J.-R., and Dehaene, S. (2014). Decoding the Dynamics of Action, Intention, and Error Detection for Conscious and Subliminal Stimuli. *The Journal of Neuroscience*, 34(4):1158.

[Charman et al., 2011] Charman, T., Pickles, A., Simonoff, E., Chandler, S., Loucas, T., and Baird, G. (2011). IQ in children with autism spectrum disorders: data from the Special Needs and Autism Project (SNAP). *Psychological Medicine*, 41(3):619–627.

[Chawarska et al., 2003] Chawarska, K., Klin, A., and Volkmar, F. (2003). Automatic Attention Cueing Through Eye Movement in 2-Year-Old Children With Autism. *Child Development*, 74(4):1108–1122. Publisher: John Wiley & Sons, Ltd.

[Chawarska and Shic, 2009] Chawarska, K. and Shic, F. (2009). Looking but not seeing: atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of autism and developmental disorders*, 39(12):1663–1672.

[Chawarska and Volkmar, 2007] Chawarska, K. and Volkmar, F. (2007). Impairments in monkey and human face recognition in 2-year-old toddlers with Autism Spectrum Disorder and Developmental Delay. *Developmental science*, 10(2):266–279.

[Chevallier et al., 2012] Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., and Schultz, R. T. (2012). The social motivation theory of autism. *Trends in cognitive sciences*, 16(4):231–239.

[Cho and Linderman, 2019] Cho, Y. S. and Linderman, K. (2019). Metacognition-based process improvement practices. *International Journal of Production Economics*, 211:132–144.

[Choi et al., 2003] Choi, I., Dalal, R., Kim-Prieto, C., and Park, H. (2003). Culture and Judgment of Causal Relevance. *Journal of personality and social psychology*, 84:46–59.

[Choi et al., 2007] Choi, I., Koo, M., and Jong An Choi (2007). Individual Differences in Analytic Versus Holistic Thinking. *Personality and Social Psychology Bulletin*, 33(5):691–705. Publisher: SAGE Publications Inc.

[Christoff and Gabrieli, 2000] Christoff, K. and Gabrieli, J. D. E. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, 28(2):168–186.

[Chung et al., 2014] Chung, Y. S., Barch, D., and Strube, M. (2014). A Meta-Analysis of Mentalizing Impairments in Adults With Schizophrenia and Autism Spectrum Disorder. *Schizophrenia Bulletin*, 40(3):602–616.

[Cialdini and Goldstein, 2004] Cialdini, R. B. and Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, 55(1):591–621.

[Cleeremans et al., 2020] Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., Vuillaume, L., and de Heering, A. (2020). Learning to Be Conscious. *Trends in cognitive sciences*, 24(2):112–123.

[Condon and Revelle, 2014] Condon, D. M. and Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, 43:52–64.

[Condon and Revelle, 2016] Condon, D. M. and Revelle, W. (2016). Selected ICAR Data from the SAPA-Project: Development and Initial Validation of a Public-Domain Measure. *Journal of Open Psychology Data*, 4(1):1.

[Cripps, 1998] Cripps, M. J. (1998). Modernization and Postmodernization: Cultural, Economic and Political Change. *The Review of Politics*, 60(2):396+.

[David et al., 2008] David, N., Gawronski, A., Santos, N. S., Huff, W., Lehnhardt, F.-G., Newen, A., and Vogeley, K. (2008). Dissociation Between Key Processes of Social Cognition in Autism: Impaired Mentalizing But Intact Sense of Agency. *Journal of Autism and Developmental Disorders*, 38(4):593–605.

[D'Cruz et al., 2016] D'Cruz, A.-M., Mosconi, M. W., Ragozzino, M. E., Cook, E. H., and Sweeney, J. A. (2016). Alterations in the functional neural circuitry supporting flexible choice behavior in autism spectrum disorders. *Translational Psychiatry*, 6(10):e916–e916.

[De Martino et al., 2017] De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., and Love, B. C. (2017). Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability. *The Journal of Neuroscience*, 37(25):6066–6074.

[De Martino et al., 2013] De Martino, B., Fleming, S. M., Garrett, N., and Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, 16(1):105–110.

[Dehaene et al., 1994] Dehaene, S., Posner, M. I., and Tucker, D. M. (1994). Localization of a Neural System for Error Detection and Compensation. *Psychological Science*, 5(5):303–305.

[Dehaene-Lambertz and Spelke, 2015] Dehaene-Lambertz, G. and Spelke, E. S. (2015). The Infancy of the Human Brain. *Neuron*, 88(1):93–109.

[Descartes, 1644] Descartes, R. (1644). *Principia Philosophiae*. Apud Danielem Elzevirium.

[Dimaggio and Lysaker, 2015] Dimaggio, G. and Lysaker, P. H. (2015). Metacognition and Mentalizing in the Psychotherapy of Patients With Psychosis and Personality Disorders. *Journal of Clinical Psychology*, 71(2):117–124.

[Doyle, 2020] Doyle, N. (2020). Neurodiversity at work: a biopsychosocial model and the impact on working adults. *British Medical Bulletin*, 135(1):108–125.

[Dunlosky and Metcalfe, 2009] Dunlosky, J. and Metcalfe, J. (2009). *Metacognition.* Metacognition. Sage Publications, Inc, Thousand Oaks, CA, US. Pages: ix, 334.

[Dunn et al., 1991] Dunn, J., Brown, J., and Beardsall, L. (1991). Family talk about feeling states and children's later understanding of others' emotions. *Developmental Psychology*, 27(3):448–455. Place: US Publisher: American Psychological Association.

[Dunn and Brown, 1993] Dunn, J. and Brown, J. R. (1993). Early conversations about causality: Content, pragmatics and developmental change. *British Journal of Developmental Psychology*, 11(2):107–123. Place: United Kingdom Publisher: British Psychological Society.

[Elsabbagh et al., 2013] Elsabbagh, M., Gliga, T., Pickles, A., Hudry, K., Charman, T., and Johnson, M. H. (2013). The development of face orienting mechanisms in infants at-risk for autism. *SI:Neurobiology of Autism*, 251:147–154.

[Engström et al., 2003] Engström, I., Ekström, L., and Emilsson, B. (2003). Psychosocial Functioning in a Group of Swedish Adults with Asperger Syndrome or High- Functioning Autism. *Autism*, page 12.

[Eriksson et al., 2013] Eriksson, J. M., Andersen, L. M., and Bejerot, S. (2013). RAADS-14 Screen: validity of a screening tool for autism spectrum disorder in an adult psychiatric population. *Molecular Autism*, 4(1):49.

[Eriksson, 2013] Eriksson, K. (2013). Autism-spectrum traits predict humor styles in the general population. *Humor*, 26(3).

[Ewbank et al., 2016] Ewbank, M. P., von dem Hagen, E. A., Powell, T. E., Henson, R. N., and Calder, A. J. (2016). The effect of perceptual expectation on repetition suppression to faces is not modulated by variation in autistic traits. *Cortex*, 80:51–60.

[Fandakova et al., 2017] Fandakova, Y., Selmeczy, D., Leckey, S., Grimm, K. J., Wendelken, C., Bunge, S. A., and Ghetti, S. (2017). Changes in ventromedial prefrontal and insular cortex support the development of metamemory from childhood into adolescence. *Proceedings of the National Academy of Sciences*, 114(29):7582.

[Flavell, 1977] Flavell, J. H. (1977). The development of knowledge about visual perception. *Nebraska Symposium on Motivation*, 25:43–76. Place: US Publisher: University of Nebraska Press.

[Flavell, 1979] Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10):906–911. Place: US Publisher: American Psychological Association.

[Flavell, 2000] Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development*, 24(1):15–23. Place: United Kingdom Publisher: Taylor & Francis.

[Fleming, 2017] Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, 2017(1).

[Fleming and Daw, 2017] Fleming, S. M. and Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, 124(1):91–114.

[Fleming and Dolan, 2012] Fleming, S. M. and Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349. _eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2011.0417.

[Fleming and Lau, 2014] Fleming, S. M. and Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8.

[Fleming et al., 2018] Fleming, S. M., van der Putten, E. J., and Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature neuroscience*, 21(4):617–624.

[Fleming et al., 2010] Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., and Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science (New York, N.Y.)*, 329(5998):1541–1543.

[Folke et al., 2016] Folke, T., Jacobsen, C., Fleming, S. M., and De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1):0002.

[Friston et al., 2013] Friston, K. J., Lawson, R., and Frith, C. D. (2013). On hyperpriors and hypopriors: comment on Pellicano and Burr. *Trends in Cognitive Sciences*, 17(1):1.

[Frith, 2012] Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2213–2223.

[Frith and Frith, 2012] Frith, C. D. and Frith, U. (2012). Mechanisms of Social Cognition. *Annual Review of Psychology*, 63(1):287–313.

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

[Frith and Happé, 1999] Frith, U. and Happé, F. (1999). Theory of Mind and Self-Consciousness: What Is It Like to Be Autistic? *Mind & Language*, 14(1):82–89. Publisher: John Wiley & Sons, Ltd.

[Fusaroli et al., 2012] Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., and Tylén, K. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, 23(8):931–939. Publisher: SAGE Publications Inc.

[Galvin et al., 2003] Galvin, S. J., Podd, J. V., Drga, V., and Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, 10(4):843–876.

[Gazzaniga, 1983] Gazzaniga, M. S. (1983). Right hemisphere language following brain bisection. A 20-year perspective. *The American psychologist*, 38(5):525–537. Place: United States.

[Gazzaniga, 1995] Gazzaniga, M. S. (1995). Principles of human brain organization derived from split-brain studies. *Neuron*, 14(2):217–228.

[Gazzaniga, 2000] Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, 123(7):1293–1326.

[Gazzaniga et al., 1962] Gazzaniga, M. S., Bogen, J. E., and Sperry, R. W. (1962). Some functional effects of sectioning the cerebral commissures in man. *Proceedings of the National Academy of Sciences*, 48(10):1765.

[Gazzaniga et al., 1977] Gazzaniga, M. S., LeDoux, J. E., and Wilson, D. H. (1977). Language, praxis, and the right hemisphere: Clues to some mechanisms of consciousness. *Neurology*, 27(12):1144–1147. Place: US Publisher: Lippincott Williams & Wilkins.

[Geurts et al., 2020] Geurts, H. M., Pol, S. E., Lobbestael, J., and Simons, C. J. P. (2020). Executive Functioning in 60+ Autistic Males: The Discrepancy Between Experienced

Challenges and Cognitive Performance. *Journal of Autism and Developmental Disorders*, 50(4):1380–1390.

[Gherman and Philiastides, 2018] Gherman, S. and Philiastides, M. G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLife*, 7:e38293. Publisher: eLife Sciences Publications, Ltd.

[Gold and Shadlen, 2007] Gold, J. I. and Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, 30(1):535–574.

[Goldman and de Vignemont, 2009] Goldman, A. and de Vignemont, F. (2009). Is social cognition embodied? *Trends in cognitive sciences*, 13(4):154–159.

[Goldman, 1993] Goldman, A. I. (1993). The psychology of folk psychology. *Behavioral and Brain Sciences*, 16(1):15–28.

[Gomez-Beldarrain et al., 2004] Gomez-Beldarrain, M., Harries, C., Garcia-Monco, J. C., Ballus, E., and Grafman, J. (2004). Patients with right frontal lesions are unable to assess and use advice to make predictive judgments. *Journal of cognitive neuroscience*, 16(1):74–89.

[Gopnik, 1993] Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16(1):1–14, 29–113. Place: United Kingdom Publisher: Cambridge University Press.

[Goupil and Kouider, 2016] Goupil, L. and Kouider, S. (2016). Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants. *Current Biology*, 26(22):3038–3045. Publisher: Elsevier.

[Goupil and Kouider, 2019] Goupil, L. and Kouider, S. (2019). Developing a Reflective Mind: From Core Metacognition to Explicit Self-Reflection. *Current Directions in Psychological Science*, 28(4):403–408.

[Grainger et al., 2016a] Grainger, C., Williams, D. M., and Lind, S. E. (2016a). Judgment of Learning Accuracy in High-functioning Adolescents and Adults with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 46(11):3570–3582.

[Grainger et al., 2016b] Grainger, C., Williams, D. M., and Lind, S. E. (2016b). Metacognitive monitoring and control processes in children with autism spectrum disorder: Diminished judgement of confidence accuracy. *Consciousness and Cognition*, 42:65–74.

[Green and Swets, 1966] Green, D. M. and Swets, J. A. (1966). Signal detection theory and psychophysics. *Signal detection theory and psychophysics.*, pages xi, 455–xi, 455.

[Greene et al., 2019] Greene, R. K., Zheng, S., Kinard, J. L., Mosner, M. G., Wiesen, C. A., Kennedy, D. P., and Dichter, G. S. (2019). Social and nonsocial visual prediction errors in autism spectrum disorder. *Autism Research*, 12(6):878–883.

[Greimel et al., 2013] Greimel, E., Bartling, J., Dunkel, J., Brückl, M., Deimel, W., Remschmidt, H., Kamp-Becker, I., and Schulte-Körne, G. (2013). The temporal dynamics of coherent motion processing in autism spectrum disorder: Evidence for a deficit in the dorsal pathway. *Behavioural Brain Research*, 251:168–175.

[Hampton, 2008] Hampton (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, 105(18):6741–6746.

[Happé, 2003] Happé, F. (2003). Theory of Mind and the Self. *Annals of the New York Academy of Sciences*, 1001(1):134–144. Publisher: John Wiley & Sons, Ltd.

[Happé and Frith, 2006] Happé, F. and Frith, U. (2006). The Weak Coherence Account: Detail-focused Cognitive Style in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 36(1):5–25.

[Happé, 1994] Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2):129–154.

[Harrison et al., 2020] Harrison, O. K., Garfinkel, S. N., Marlow, L., Finnegan, S., Marino, S., Nanz, L., Allen, M., Finnemann, J., Keur-Huizinga, L., Harrison, S. J., Stephan, K. E., Pattinson, K., and Fleming, S. M. (2020). The Filter Detection Task for measurement of breathing-related interoception and metacognition. *bioRxiv*, page 2020.06.29.176941.

[Harvey and Fischer, 1997] Harvey, N. and Fischer, I. (1997). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes*, 70(2):117–133.

[Helles et al., 2017] Helles, A., Gillberg, I. C., Gillberg, C., and Billstedt, E. (2017). Asperger syndrome in males over two decades: Quality of life in relation to diagnostic stability and psychiatric comorbidity. *Autism*, 21(4):458–469.

[Hembacher and Ghetti, 2014] Hembacher, E. and Ghetti, S. (2014). Don't Look at My Answer: Subjective Uncertainty Underlies Preschoolers' Exclusion of Their Least Accurate Memories. *Psychological Science*, 25(9):1768–1776. Publisher: SAGE Publications Inc.

[Hertz et al., 2017] Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., and Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications*, 8(1):2191.

[Herzallah et al., 2013] Herzallah, M., Moustafa, A., Natsheh, J., Abdellatif, S., Taha, M., Tayem, Y., Sehwail, M., Amleh, I., Petrides, G., Myers, C., and Gluck, M. (2013). Learning from negative feedback in patients with major depressive disorder is attenuated by SSRI antidepressants. *Frontiers in Integrative Neuroscience*, 7:67.

[Heyes, 2018] Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking.* Cognitive gadgets: The cultural evolution of thinking. Harvard University Press, Cambridge, MA, US. Pages: 292.

[Heyes et al., 2020] Heyes, C., Bang, D., Shea, N., Frith, C. D., and Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences*, 24(5):349–362.

[Heyes and Frith, 2014] Heyes, C. M. and Frith, C. D. (2014). The cultural evolution of mind reading. *Science (New York, N.Y.)*, 344(6190):1243091.

[Hilgenstock et al., 2014] Hilgenstock, R., Weiss, T., and Witte, O. W. (2014). You'd Better Think Twice: Post-Decision Perceptual Confidence. *NeuroImage*, 99:323–331.

[Hofstede, 2011] Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. page 26.

[Holt et al., 2011] Holt, D. J., Cassidy, B. S., Andrews-Hanna, J. R., Lee, S. M., Coombs, G., Goff, D. C., Gabrieli, J. D., and Moran, J. M. (2011). An anterior-to-posterior shift in midline cortical activity in schizophrenia during self-reflection. *Biological psychiatry*, 69(5):415–423. Edition: 2010/12/08.

[Hooker et al., 2011] Hooker, C. I., Bruce, L., Lincoln, S. H., Fisher, M., and Vinogradov, S. (2011). Theory of Mind Skills Are Related to Gray Matter Volume in the Ventromedial Prefrontal Cortex in Schizophrenia. *Biological Psychiatry*, 70(12):1169–1178.

[Howlin et al., 2004] Howlin, P., Goode, S., Hutton, J., and Rutter, M. (2004). Adult outcome for children with autism. *Journal of Child Psychology and Psychiatry*, 45(2):212–229.

[Howlin et al., 2013] Howlin, P., Moss, P., Savage, S., and Rutter, M. (2013). Social Outcomes in Mid- to Later Adulthood Among Individuals Diagnosed With Autism and Average Nonverbal IQ as Children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 52(6):572–581.e1.

[Hu and Bentler,] Hu, L.-t. and Bentler, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1):1–55.

[Huang et al., 2007] Huang, D. T., Clermont, G., Sexton, B. J., Karlo, C. A., Miller, R. G., Weissfeld, L. A., Rowan, K. M., and Angus, D. C. (2007). Perceptions of safety culture vary across the intensive care units of a single institution. *Critical Care Medicine*, 35(1).

[Isaacson and Fujita, 2006] Isaacson, R. and Fujita, F. (2006). Metacognitive knowledge monitoring and self-regulated learning: Academic success and reflections on learning. *Journal of the Scholarship of Teaching and Learning*, 6:39–55.

[Jacob et al., 2019] Jacob, A. M., Datta, M., Kumpatla, S., Selvaraj, P., and Viswanthan, V. (2019). Prevalence of Diabetes Mellitus and Exposure to Suspended Particulate Matter. *Journal of health & pollution*, 9(22):190608.

[Jenkins and Astington, 1996] Jenkins, J. M. and Astington, J. W. (1996). Cognitive factors and family structure associated with theory of mind development in young children. *Developmental Psychology*, 32(1):70–78. Place: US Publisher: American Psychological Association.

[Ji et al., 2001] Ji, L.-J., Nisbett, R. E., and Su, Y. (2001). Culture, Change, and Prediction. *Psychological Science*, 12(6):450–456.

[Johnson and Fowler, 2011] Johnson, D. D. P. and Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, 477(7364):317–320.

[Johnson et al., 2006] Johnson, S. A., Yechiam, E., Murphy, R. R., Queller, S., and Stout, J. C. (2006). Motivational processes and autonomic responsivity in Asperger's disorder: Evidence from the Iowa Gambling Task. *Journal of the International Neuropsychological Society*, 12(5):668–676.

[Jolliffe and Baron-Cohen, 1999] Jolliffe, T. and Baron-Cohen, S. (1999). A test of central coherence theory: linguistic processing in high-functioning adults with autism or Asperger syndrome: is local coherence impaired? *Cognition*, 71(2):149–185.

[Joseph et al., 2002] Joseph, R. M., Tager-Flusberg, H., and Lord, C. (2002). Cognitive profiles and social-communicative functioning in children with autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 43(6):807–821.

[Jänsch and Hare, 2014] Jänsch, C. and Hare, D. J. (2014). An Investigation of the "Jumping to Conclusions" Data-Gathering Bias and Paranoid Thoughts in Asperger Syndrome. *Journal of Autism and Developmental Disorders*, 44(1):111–119.

[Kana et al., 2015] Kana, R. K., Maximo, J. O., Williams, D. L., Keller, T. A., Schipul, S. E., Cherkassky, V. L., Minshew, N. J., and Just, M. A. (2015). Aberrant functioning of the theory-of-mind network in children and adolescents with autism. *Molecular Autism*, 6(1):59.

[Kappes et al., 2020] Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., and Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, 23(1):130–137.

[Kelliher et al., 2011] Kelliher, C., Bobek, D., and Hageman, A. (2011). The Social Norms of Tax Compliance: Scale Development, Social Desirability and Presentation Effects. In *Advances in Accounting Behavioral Research*, volume 14.

[Kenny et al., 2016] Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., and Pellicano, E. (2016). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism*, 20(4):442–462.

[Khalvati et al., 2021] Khalvati, K., Kiani, R., and Rao, R. P. N. (2021). Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nature Communications*, 12(1):5704.

[Kiani et al., 2014] Kiani, R., Corthell, L., and Shadlen, M. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, 84(6):1329–1342.

[Kiani and Shadlen, 2009] Kiani, R. and Shadlen, M. N. (2009). Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science*, 324(5928):759.

[Kim and Shadlen, 1999] Kim, J.-N. and Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, 2(2):176–185.

[Kim et al., 2018] Kim, S., Shahaeian, A., and Proust, J. (2018). Developmental diversity in mindreading and metacognition. In *Metacognitive diversity: An interdisciplinary approach.*, pages 97–133. Oxford University Press, New York, NY, US.

[Kim et al., 2020] Kim, S., Sodian, B., Paulus, M., Senju, A., Okuno, A., Ueno, M., Itakura, S., and Proust, J. (2020). Metacognition and mindreading in young children: A cross-cultural study. *Consciousness and Cognition*, 85:103017.

[Koren et al., 2006] Koren, D., Seidman, L. J., Goldsmith, M., and Harvey, P. D. (2006). Real-World Cognitive–and Metacognitive–Dysfunction in Schizophrenia: A New Approach for Measuring (and Remediating) More ”Right Stuff”. *Schizophrenia Bulletin*, 32(2):310–326.

[Korn et al., 2014] Korn, C., Fan, Y., Zhang, K., Wang, C., Han, S., and Heekeren, H. (2014). Cultural influences on social feedback processing of character traits. *Frontiers in Human Neuroscience*, 8:192.

[Kruschke, 2010] Kruschke, J. K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, Inc., USA, 1st edition.

[Kuzmanovic et al., 2019] Kuzmanovic, B., Rigoux, L., and Vogeley, K. (2019). Brief Report: Reduced Optimism Bias in Self-Referential Belief Updating in High-Functioning Autism. *Journal of Autism and Developmental Disorders*, 49(7):2990–2998.

[Lamash and Josman, 2021] Lamash, L. and Josman, N. (2021). A metacognitive intervention model to promote independence among individuals with autism spectrum disorder: Implementation on a shopping task in the community. *Neuropsychological Rehabilitation*, 31(2):189–210. Publisher: Routledge.

[Lana et al., 2017] Lana, F., Cruz, M., Victor, P., and Martí-Bonany, J. (2017). Social Cognition Based Therapies for People with Schizophrenia: Focus on Metacognitive and Mentalization Approaches. *Schizophrenia Research*, 1(15).

[Large et al., 2019] Large, I., Pellicano, E., Mojzisch, A., and Krug, K. (2019). Developmental trajectory of social influence integration into perceptual decisions in children. *Proceedings of the National Academy of Sciences*, 116(7):2713–2722.

[Larson et al., 2011] Larson, M. J., South, M., Krauskopf, E., Clawson, A., and Crowley, M. J. (2011). Feedback and reward processing in high-functioning autism. *Psychiatry Research*, 187(1-2):198–203.

[Leslie, 1987] Leslie, A. M. (1987). Pretense and representation: The origins of ”theory of mind.”. *Psychological Review*, 94(4):412–426.

[Leslie and Frith, 1988] Leslie, A. M. and Frith, U. (1988). Autistic children's understanding of seeing, knowing and believing. *British Journal of Developmental Psychology*, 6(4):315–324. _eprint: https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-835X.1988.tb01104.x.

[Lieder et al., 2019] Lieder, I., Adam, V., Frenkel, O., Jaffe-Dax, S., Sahani, M., and Ahissar, M. (2019). Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nature Neuroscience*, 22(2):256–264.

[Liu et al., 2008] Liu, D., Wellman, H. M., Tardif, T., and Sabbagh, M. A. (2008). Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology*, 44(2):523–531. Place: US Publisher: American Psychological Association.

[Livingston et al., 2019a] Livingston, L. A., Colvert, E., the Social Relationships Study Team, Bolton, P., and Happé, F. (2019a). Good social skills despite poor theory of mind: exploring compensation in autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 60(1):102–110.

[Livingston et al., 2019b] Livingston, L. A., Shah, P., and Happé, F. (2019b). Compensatory strategies below the behavioural surface in autism: a qualitative study. *The Lancet Psychiatry*, 6(9):766–777.

[Livingston et al., 2021] Livingston, L. A., Shah, P., White, S. J., and Happé, F. (2021). Further developing the frith–happé animations: A quicker, more objective, and web-based test of theory of mind for autistic and neurotypical adults. *Autism Research*, 14(9):1905–1912. Publisher: John Wiley & Sons, Ltd.

[Lockl and Schneider, 2007] Lockl, K. and Schneider, W. (2007). Knowledge about the mind: links between theory of mind and later metamemory. *Child development*, 78(1):148–167. Place: United States.

[Loomes et al., 2017] Loomes, R., Hull, L., and Mandy, W. P. L. (2017). What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6):466–474.

[Luce, 1986] Luce (1986). *Response Times*.

[Luke et al., 2012] Luke, L., Clare, I. C., Ring, H., Redley, M., and Watson, P. (2012). Decision-making difficulties experienced by adults with autism spectrum conditions. *Autism*, 16(6):612–621.

[Luman et al., 2009] Luman, M., Van Meel, C. S., Oosterlaan, J., Sergeant, J. A., and Geurts, H. M. (2009). Does reward frequency or magnitude drive reinforcement-learning in attention-deficit/hyperactivity disorder? *Psychiatry Research*, 168(3):222–229.

[Lysaker et al., 2012] Lysaker, P. H., Vohs, J. L., Ballard, R., Fogley, R., Salvatore, G., Popolo, R., and Dimaggio, G. (2012). Metacognition, self-reflection and recovery in schizophrenia. *Future Neurology*, 8(1):103–115. Publisher: Future Medicine.

[Mahmoodi et al., 2015] Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C. D., Roepstorff, A., Rees, G., and Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, 112(12):3835–3840.

[Maniscalco and Lau, 2012] Maniscalco, B. and Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1):422–430.

[Maniscalco and Lau, 2014] Maniscalco, B. and Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance SDT model. *The cognitive neuroscience of metacognition.*, pages 25–66. Place: New York, NY, US Publisher: Springer-Verlag Publishing.

[Maras et al., 2017] Maras, K., Gamble, T., and Brosnan, M. (2017). Supporting metacognitive monitoring in mathematics learning for young people with autism spectrum disorder: A classroom-based study. *Autism*, 23(1):60–70. Publisher: SAGE Publications Ltd.

[Markus and Kitayama, 2010] Markus, H. R. and Kitayama, S. (2010). Cultures and Selves: A Cycle of Mutual Constitution. *Perspectives on Psychological Science*, 5(4):420–430. Publisher: SAGE Publications Inc.

[Marshall-Mies et al., 2000] Marshall-Mies, J. C., Fleishman, E. A., Martin, J. A., Zaccaro, S. J., Baughman, W. A., and McGee, M. L. (2000). Development and evaluation of cognitive and metacognitive measures for predicting leadership potential. *The Leadership Quarterly*, 11(1):135–153.

[Martino et al., 2007] Martino, D. J., Bucay, D., Butman, J. T., and Allegri, R. F. (2007). Neuropsychological frontal impairments and negative symptoms in schizophrenia. *Psychiatry Research*, 152(2):121–128.

[Masson and Rotello, 2009] Masson, M. E. J. and Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. *Journal of experimental psychology. Learning, memory, and cognition*, 35(2):509–527. Place: United States.

[Mayer and Träuble, 2015] Mayer, A. and Träuble, B. (2015). The Weird World of Cross-Cultural False-Belief Research: A True- and False-Belief Study Among Samoan Children Based on Commands. *Journal of Cognition and Development*, 16(4):650–665.

[Mayer and Träuble, 2013] Mayer, A. and Träuble, B. E. (2013). Synchrony in the onset of mental state understanding across cultures? A study among children in samoa. *International Journal of Behavioral Development*, 37(1):21–28.

[McGauley et al., 2011] McGauley, G., Yakeley, J., Williams, A., and Bateman, A. (2011). Attachment, mentalization and antisocial personality disorder: The possible contribution of mentalization-based treatment. *European Journal of Psychotherapy & Counselling*, 13(4):371–393. Publisher: Routledge.

[McMahon et al., 2016] McMahon, C. M., Henderson, H. A., Newell, L., Jaime, M., and Mundy, P. (2016). Metacognitive Awareness of Facial Affect in Higher-Functioning Children and Adolescents with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 46(3):882–898.

[McPartland et al., 2012] McPartland, J. C., Crowley, M. J., Perszyk, D. R., Mukerji, C. E., Naples, A. J., Wu, J., and Mayes, L. C. (2012). Preserved reward outcome processing in ASD as revealed by event-related potentials. *Journal of Neurodevelopmental Disorders*, 4(1):16.

[Mesoudi et al., 2015] Mesoudi, A., Chang, L., Murray, K., and Lu, H. J. (2015). Higher frequency of social learning in China than in the West shows cultural variation in the dynamics of cultural evolution. *Proceedings of the Royal Society B: Biological Sciences*, 282(1798):20142209.

[Milgram, 1963] Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371–378.

[Milne et al., 2002] Milne, E., Swettenham, J., Hansen, P., Campbell, R., Jeffries, H., and Plaisted, K. (2002). High motion coherence thresholds in children with autism. *Journal of Child Psychology and Psychiatry*, 43(2):255–263.

[Modinos et al., 2011] Modinos, G., Renken, R., Ormel, J., and Aleman, A. (2011). Self-reflection and the psychosis-prone brain: an fMRI study. *Neuropsychology*, 25(3):295–305. Place: United States.

[Moore et al., 2018] Moore, D. A., Dev, A. S., and Goncharova, E. Y. (2018). Overconfidence Across Cultures. *Collabra: Psychology*, 4(36).

[Moss et al., 2017] Moss, P., Mandy, W., and Howlin, P. (2017). Child and Adult Factors Related to Quality of Life in Adults with Autism. *Journal of Autism and Developmental Disorders*, 47(6):1830–1837.

[Murphy et al., 2015] Murphy, P. R., Robertson, I. H., Harty, S., and O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife*, 4:e11946. Publisher: eLife Sciences Publications, Ltd.

[Myers et al., 2013] Myers, C. E., Moustafa, A. A., Sheynin, J., VanMeenen, K. M., Gilbertson, M. W., Orr, S. P., Beck, K. D., Pang, K. C. H., and Servatius, R. J. (2013). Learning to Obtain Reward, but Not Avoid Punishment, Is Affected by Presence of PTSD Symptoms in Male Veterans: Empirical Data and Computational Model. *PLOS ONE*, 8(8):e72508.

[Nan et al., 2006] Nan, Y., Knösche, T. R., and Friederici, A. D. (2006). The perception of musical phrase structure: A cross-cultural ERP study. *Brain Research*, 1094(1):179–191.

[Narang and Saini, 2013] Narang, D. and Saini, S. (2013). Metacognition and Academic Performance of Rural Adolescents. *Studies on Home and Community Science*, 7(3):167–175.

[Navajas et al., 2016] Navajas, J., Bahrami, B., and Latham, P. E. (2016). Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences*, 11:55–60.

[Nelson, 1984] Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1):109–133. Place: US Publisher: American Psychological Association.

[Nelson, 1990] Nelson, T. O. (1990). Metamemory: A Theoretical Framework and New Findings. volume 26 of *Psychology of Learning and Motivation*, pages 125–173. Academic Press. ISSN: 0079-7421.

[Nichols and Stich, 2003] Nichols, S. and Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Mindreading: An integrated account of pretence, self-awareness, and understanding other minds. Clarendon Press/Oxford University Press, New York, NY, US. Pages: 237.

[Nicholson et al., 2020] Nicholson, T., Williams, D., Lind, S., Grainger, C., and Carruthers, P. (2020). Linking metacognition and mindreading: Evidence from autism and dual-task investigations. *Journal of experimental psychology. General*.

[Nicholson et al., 2019] Nicholson, T., Williams, D. M., Grainger, C., Lind, S. E., and Carruthers, P. (2019). Relationships between implicit and explicit uncertainty monitoring and mindreading: Evidence from autism spectrum disorder. *Consciousness and Cognition*, 70:11–24.

[Oeberst and Wu, 2015] Oeberst, A. and Wu, S. (2015). Independent vs. Interdependent self-construal and interrogative compliance: Intra- and cross-cultural evidence. *Personality and Individual Differences*, 85:50–55. Place: Netherlands Publisher: Elsevier Science.

[Olsen et al., 2019] Olsen, K., Roepstorff, A., and Bang, D. (2019). *Knowing whom to learn from: individual differences in metacognition and weighting of social information*.

[Oyserman, 1993] Oyserman, D. (1993). The lens of personhood: Viewing the self and others in a multicultural society. *Journal of Personality and Social Psychology*, 65(5):993–1009. Place: US Publisher: American Psychological Association.

[Palan and Schitter, 2018] Palan, S. and Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

[Palmer et al., 2017] Palmer, C. J., Lawson, R. P., and Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological bulletin*, 143(5):521–542.

[Palser et al., 2018] Palser, E. R., Fotopoulou, A., and Kilner, J. M. (2018). Altering movement parameters disrupts metacognitive accuracy. *Consciousness and Cognition: An International Journal*, 57:33–40.

[Pariser, 1981] Pariser, D. (1981). Nadia's Drawings: Theorizing about an Autistic Child's Phenomenal Ability. *Studies in Art Education*, 22(2):20–31.

[Patel et al., 2012] Patel, D., Fleming, S. M., and Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748):4853–4860.

[Peiker et al., 2015] Peiker, I., Schneider, T. R., Milne, E., Schöttle, D., Vogeley, K., Münchau, A., Schunke, O., Siegel, M., Engel, A. K., and David, N. (2015). Stronger Neural Modulation by Visual Motion Intensity in Autism Spectrum Disorders. *PLOS ONE*, 10(7):e0132531.

[Perner et al., 1994] Perner, J., Ruffman, T., and Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child Development*, 65(4):1228–1238. Place: United Kingdom Publisher: Blackwell Publishing.

[Pernet et al., 2013] Pernet, C., Wilcox, R., and Rousselet, G. (2013). Robust Correlation Analyses: False Positive and Power Validation Using a New Open Source Matlab Toolbox. *Frontiers in Psychology*, 3:606.

[Pescetelli and Yeung, 2021] Pescetelli, N. and Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General*, 150(3):507–526.

[Peterson and Siegal, 1995] Peterson, C. C. and Siegal, M. (1995). Deafness, Conversation and Theory of Mind. *Journal of Child Psychology and Psychiatry*, 36(3):459–474. _eprint: https://acamh.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-7610.1995.tb01303.x.

[Pirrone et al., 2017] Pirrone, A., Dickinson, A., Gomez, R., Stafford, T., and Milne, E. (2017). Understanding perceptual judgment in autism spectrum disorder using the drift diffusion model. *Neuropsychology*, 31(2):173–180.

[Plaisted et al.,] Plaisted, K., O'Riordan, M., and Baron-Cohen, S. Enhanced Discrimination of Novel, Highly Similar Stimuli by Adults with Autism During a Perceptual Learning Task. page 11.

[Pleskac and Busemeyer, 2010] Pleskac, T. J. and Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, 117(3):864–901.

[Pouget et al., 2016] Pouget, A., Drugowitsch, J., and Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3):366–374.

[Powell et al., 2016] Powell, G., Meredith, Z., McMillin, R., and Freeman, T. C. A. (2016). Bayesian Models of Individual Differences: Combining Autistic Traits and Sensory Thresholds to Predict Motion Perception. *Psychological Science*, 27(12):1562–1572.

[Proust and Fortier, 2018] Proust, J. and Fortier, M. (2018). *Metacognitive Diversity: An Interdisciplinary Approach*. Publication Title: Metacognitive Diversity: An Interdisciplinary Approach.

[Rabbitt, 1966] Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, 71(2):264–272. Place: US Publisher: American Psychological Association.

[Rahnev and Fleming, 2019] Rahnev, D. and Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, 2019(niz009).

[Ratcliff, 1978] Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2):59–108.

[Reed, 2019] Reed, P. (2019). Unpredictability reduces over-selective responding of individuals with ASD who have language impairments. *Research in Autism Spectrum Disorders*, 57:35–45.

[Resulaj et al., 2009] Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, 461(7261):263–266.

[Richardson et al., 2020] Richardson, H., Koster-Hale, J., Caselli, N., Magid, R., Benedict, R., Olson, H., Pyers, J., and Saxe, R. (2020). Reduced neural selectivity for mental states in deaf children with delayed exposure to sign language. *Nature Communications*, 11(1):3246.

[Ridderinkhof et al., 2004] Ridderinkhof, K. R., van den Wildenberg, W. P., Segalowitz, S. J., and Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: The role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Neurocognitive mechanisms of performance monitoring and inhibitory control*, 56(2):129–140.

[Robic et al., 2015] Robic, S., Sonié, S., Fonlupt, P., Henaff, M.-A., Touil, N., Coricelli, G., Mattout, J., and Schmitz, C. (2015). Decision-Making in a Changing World: A Study in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 45(6):1603–1613.

[Robinson et al., 2016] Robinson, E. B., St Pourcain, B., Anttila, V., Kosmicki, J. A., Bulik-Sullivan, B., Grove, J., Maller, J., Samocha, K. E., Sanders, S. J., Ripke, S., Martin, J., Hollegaard, M. V., Werge, T., Hougaard, D. M., Neale, B. M., Evans, D. M., Skuse, D., Mortensen, P. B., Børglum, A. D., Ronald, A., Smith, G. D., and Daly, M. J. (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nature genetics*, 48(5):552–555.

[Roebers, 2017] Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, 45:31–51.

[Rollwage et al., 2018] Rollwage, M., Dolan, R. J., and Fleming, S. M. (2018). Metacognitive Failure as a Feature of Those Holding Radical Beliefs. *Current Biology*, 28(24):4014–4021.e8.

[Rollwage et al., 2020] Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., and Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, 11(1):2634.

[Rosenblau et al., 2015] Rosenblau, G., Kliemann, D., Heekeren, H. R., and Dziobek, I. (2015). Approximating Implicit and Explicit Mentalizing with Two Naturalistic Video-Based Tasks in Typical Development and Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 45(4):953–965.

[Rouault et al., 2019] Rouault, M., Dayan, P., and Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, 10(1):1141.

[Rouault et al., 2018] Rouault, M., McWilliams, A., Allen, M. G., and Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*, 1:e17.

[Ryle, 1949] Ryle, G. (1949). Meaning and Necessity. *Philosophy*, 24(88):69–76. Publisher: Cambridge University Press.

[Sapey-Triomphe et al., 2018] Sapey-Triomphe, L.-A., Sonié, S., Hénaff, M.-A., Mattout, J., and Schmitz, C. (2018). Adults with Autism Tend to Undermine the Hidden Environmental Structure: Evidence from a Visual Associative Learning Task. *Journal of Autism and Developmental Disorders*, 48(9):3061–3074.

[Schulz et al., 2020] Schulz, L., Rollwage, M., Dolan, R. J., and Fleming, S. M. (2020). Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences*, 117(49):31527.

[Shah et al., 2016] Shah, P., Catmur, C., and Bird, G. (2016). Emotional decision-making in autism spectrum disorder: the roles of interoception and alexithymia. *Molecular Autism*, 7(1):43.

[Shea et al., 2014] Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., and Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in cognitive sciences*, 18(4):186–193.

[Shields-Wolfe and Gallagher, 1992] Shields-Wolfe, J. and Gallagher, P. A. (1992). Functional utilization of splinter skills for the employment of a young adult with autism. *Focus on Autistic Behavior*, 7(4):1–16.

[Shimamura and Squire, 1991] Shimamura, A. P. and Squire, L. R. (1991). The relationship between fact and source memory: Findings from amnesic patients and normal subjects. *Psychobiology*, 19(1):1–10.

[Singelis, 1994] Singelis, T. M. (1994). The Measurement of Independent and Interdependent Self-Construals. *Personality and Social Psychology Bulletin*, 20(5):580–591.

[Smith and Ratcliff, 2004] Smith, P. L. and Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, 27(3):161–168.

[Sniezek and Van Swol, 2001] Sniezek, J. A. and Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge–advisor system. *Organizational Behavior and Human Decision Processes*, 84(2):288–307.

[Sodian and Frith, 1992] Sodian, B. and Frith, U. (1992). Deception and sabotage in autistic, retarded and normal children. *Child Psychology & Psychiatry & Allied Disciplines*, 33(3):591–605. Place: United Kingdom Publisher: Pergamon Press.

[Solomon et al., 2015] Solomon, M., Frank, M. J., Ragland, J. D., Smith, A. C., Niendam, T. A., Lesh, T. A., Grayson, D. S., Beck, J. S., Matter, J. C., and Carter, C. S. (2015). Feedback-Driven Trial-by-Trial Learning in Autism Spectrum Disorders. *American Journal of Psychiatry*, 172(2):173–181.

[Spek et al., 2009] Spek, A., Schatorjé, T., Scholte, E., and van Berckelaer-Onnes, I. (2009). Verbal fluency in adults with high functioning autism or Asperger syndrome. *Neuropsychologia*, 47(3):652–656.

[Spek and Wouters, 2010] Spek, A. A. and Wouters, S. G. M. (2010). Autism and schizophrenia in high functioning adults: Behavioral differences and overlap. *Research in Autism Spectrum Disorders*, 4(4):709–717.

[St Pourcain et al., 2018] St Pourcain, B., Robinson, E. B., Anttila, V., Sullivan, B. B., Maller, J., Golding, J., Skuse, D., Ring, S., Evans, D. M., Zammit, S., Fisher, S. E., Neale, B. M., Anney, R. J. L., Ripke, S., Hollegaard, M. V., Werge, T., Ronald, A., Grove, J., Hougaard, D. M., Børglum, A. D., Mortensen, P. B., Daly, M. J., and Davey Smith, G. (2018). ASD and schizophrenia show distinct developmental profiles in common genetic overlap with population-based social communication difficulties. *Molecular psychiatry*, 23(2):263–270.

[St Pourcain et al., 2014] St Pourcain, B., Skuse, D. H., Mandy, W. P., Wang, K., Hakonarson, H., Timpson, N. J., Evans, D. M., Kemp, J. P., Ring, S. M., McArdle, W. L., Golding, J., and Smith, G. D. (2014). Variability in the common genetic architecture of social-communication spectrum phenotypes during childhood and adolescence. *Molecular autism*, 5(1):18.

[Stankov and Lee, 2014] Stankov, L. and Lee, J. (2014). Overconfidence Across World Regions. *Journal of Cross-Cultural Psychology*, 45(5):821–837.

[Sternberg, 1998] Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student? *Instructional Science*, 26(1/2):127–140. Publisher: Springer.

[Talluri et al., 2018] Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., and Donner, T. H. (2018). Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Current Biology*, 28(19):3128–3135.e8.

[Tenney et al., 2019] Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., and Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, 116(3):396–415.

[Tiokhin et al., 2021] Tiokhin, L., Hackman, J., Munira, S., Jesmin, K., and Hruschka, D. (2021). Generalizability is not optional: insights from a cross-cultural study of social discounting. *Royal Society Open Science*, 6(2):181386.

[Vaccaro and Fleming, 2018] Vaccaro, A. G. and Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances*, 2:2398212818810591.

[Valenzuela et al., 2010] Valenzuela, A., Mellers, B., and Strebel, J. (2010). Pleasurable Surprises: A Cross-Cultural Study of Consumer Responses to Unexpected Incentives. *Journal of Consumer Research*, 36(5):792–805.

[van den Berg et al., 2016] van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., and Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, 5:e12192. Publisher: eLife Sciences Publications, Ltd.

[van der Plas et al., 2019] van der Plas, E., David, A. S., and Fleming, S. M. (2019). Advice-taking as a bridge between decision neuroscience and mental capacity. *International Journal of Law and Psychiatry*, 67:101504.

[van der Plas et al., 2022] van der Plas, E., S., Z., K, D., Bang, D., Wright, N., J., L., and S., F. (2022). Isolating cultural contributors to confidence. *Journal of Experimental Psychology: General*, 2022.

[Vickers, 1979] Vickers, D. (1979). *Decision Processes in Visual Perception.*

[Wald, 1947] Wald, A. (1947). *Sequential analysis.* Sequential analysis. John Wiley, Oxford, England. Pages: xii, 212.

[Warrier et al., 2019] Warrier, V., Toro, R., Won, H., Leblond, C. S., Cliquet, F., Delorme, R., De Witte, W., Bralten, J., Chakrabarti, B., Børglum, A. D., Grove, J., Poelmans, G., Hinds, D. A., Bourgeron, T., and Baron-Cohen, S. (2019). Social and non-social autism symptoms and trait domains are genetically dissociable. *Communications Biology*, 2(1):328.

[Weber, 1905] Weber, M. (1905). *The Protestant ethic and the spirit of capitalism.* New York.

[Wegner, 2002] Wegner, D. M. (2002). *The illusion of conscious will.* The illusion of conscious will. MIT Press, Cambridge, MA, US. Pages: xi, 405.

[Weil et al., 2013] Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., and Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and cognition*, 22(1):264–271.

[Wellman and Liu, 2004] Wellman, H. M. and Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development*, 75(2):523–541.

[White et al., 2009] White, S., Hill, E., Happé, F., and Frith, U. (2009). Revisiting the Strange Stories: Revealing Mentalizing Impairments in Autism. *Child Development*, 80(4):1097–1117.

[White et al., 2011] White, S. J., Coniston, D., Rogers, R., and Frith, U. (2011). Developing the Frith-Happé animations: A quick and objective test of Theory of Mind for adults with autism. *Autism Research*, 4(2):149–154.

[Williams et al., 2018] Williams, D. M., Bergström, Z., and Grainger, C. (2018). Metacognitive monitoring and the hypercorrection effect in autism and the general population: Relation to autism(-like) traits and mindreading. *Autism*, 22(3):259–270.

[Wilson and Bishop, 2020] Wilson, A. C. and Bishop, D. V. M. (2020). Judging meaning: A domain-level difference between autistic and non-autistic adults. *Royal Society Open Science*, 7(11):200845.

[Wilson, 2002] Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636.

[Wimmer and Perner, 1983] Wimmer, H. and 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.

[Wojcik et al., 2013] Wojcik, D. Z., Moulin, C. J. A., and Souchay, C. (2013). Metamemory in children with autism: Exploring "feeling-of-knowing" in episodic and semantic memory. *Neuropsychology*, 27(1):19–27.

[Woodcock et al., 2020] Woodcock, K. A., Cheung, C., González Marx, D., and Mandy, W. (2020). Social Decision Making in Autistic Adolescents: The Role of Theory of Mind, Executive Functioning and Emotion Regulation. *Journal of Autism and Developmental Disorders*, 50(7):2501–2512.

[Wright et al., 2018] Wright, N. D., Grohn, J., Song, C., Rees, G., and Lawson, R. P. (2018). Cultural effects on computational metrics of spatial and temporal context. *Scientific Reports*, 8(1):2027.

[Yates et al., 1998] Yates, J., Lee, J.-W., Shinotsuka, H., Patalano, A. L., and Sieck, W. R. (1998). Cross-Cultural Variations in Probability Judgment Accuracy: Beyond General Knowledge Overconfidence? *Organizational Behavior and Human Decision Processes*, 74(2):89–117.

[Yates et al., 1989] Yates, J., Zhu, Y., Ronis, D. L., Wang, D.-F., Shinotsuka, H., and Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes*, 43(2):145–171.

[Yeung and Summerfield, 2012] Yeung, N. and Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1594):1310–1321.

[Yirmiya et al., 1996a] Yirmiya, N., Solomonica-Levi, D., and Shulman, C. (1996a). The ability to manipulate behavior and to understand manipulation of beliefs: A comparison of individuals with autism, mental retardation, and normal development. *Developmental Psychology*, 32(1):62–69.

[Yirmiya et al., 1996b] Yirmiya, N., Solomonica-Levi, D., Shulman, C., and Pilowsky, T. (1996b). Theory of Mind Abilities in Individuals With Autism, Down Syndrome, and

Mental Retardation of Unknown Etiology: The Role of Age and Intelligence. *Journal of child psychology and psychiatry, and allied disciplines*, 37:1003–14.

[Zalla et al., 2015] Zalla, T., Miele, D., Leboyer, M., and Metcalfe, J. (2015). Metacognition of agency and theory of mind in adults with high functioning autism. *Consciousness and Cognition*, 31:126–138.

[Zwart et al., 2018] Zwart, F. S., Vissers, C. T. W. M., and Maes, J. H. R. (2018). The Association Between Sequence Learning on the Serial Reaction Time Task and Social Impairments in Autism. *Journal of Autism and Developmental Disorders*, 48(8):2692–2700.