Metacognition facilitates Theory of Mind through optimal weighting of trait inferences

Emily L. Long*[1], Caroline Catmur[2], Stephen M. Fleming[3,4,5], Geoffrey Bird[1]

[1]Department of Experimental Psychology, University of Oxford, Oxford, OX2 6GG, UK.
[2]Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK.
[3]Wellcome Centre for Human Neuroimaging, University College London, London, WC1N 3BG, UK
[4]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, WC1B 5EH, UK
[5]Department of Experimental Psychology, University College London, London, WC1H 0AP, UK

Corresponding author: Emily L. Long, emily.long@psy.ox.ac.uk, Department of Experimental Psychology, Anna Watts Building, Woodstock Rd, Oxford OX2 6GG, UK

Metacognition facilitates Theory of Mind through optimal weighting of trait inferences

Abstract

The ability to represent and infer accurately others' mental states, known as Theory of Mind (ToM), has been theorised to be associated with metacognitive ability. Here, we considered the role of metacognition in mental state inference through the lens of a recent theoretical approach to explaining ToM, the 'Mind-space' framework. The Mind-space framework posits that trait inference, representation of the qualities of the mind giving rise to the mental state, is important in forming accurate mental state inferences. We tested a potential role for metacognition in facilitating optimal weighting of trait inferences, as well as several theoretical predictions regarding factors associated with the accuracy of trait inference and confidence in those trait inferences. Participants completed a judgement-of-confidence task in the trait inference domain alongside the Interview Task, a recently-developed task for assessing the accuracy of trait and mental state inferences. A simple relationship in which increased metacognitive sensitivity is associated with increased accuracy of mental states inferences was not found. However, when predicting trial-level performance, confidence in trait inference was shown to modulate the effect of trait inference accuracy on mental state inference accuracy. This effect was greater in magnitude with lower metacognitive sensitivity, i.e., when confidence is more likely to be misplaced. Furthermore, participants' trait inference ability was associated with the accuracy of their understanding of the average mind. In addition, the accuracy of specific trait inferences was predicted by the participant's similarity to the target, but this similarity benefit was reduced in participants whose self-perception was inaccurate. Reported confidence in a given trait inference was also predicted by participant-target similarity, such that participants showed greater overconfidence in judgements made about similar targets. This overconfidence effect was larger when self-perception was more erroneous. Results support several theoretical claims made by the Mind-space theory, and further elucidate the processes underlying accurate mental state inference.

Key Words

Theory of Mind; metacognition; individual differences; Mind-space; Interview Task; mental states.

1. Introduction

Theory of Mind (ToM), classically defined as the ability to represent the mental states of oneself and others (Premack & Woodruff, 1978), is an important feature of human social cognition. Although definitions vary, here a mental state is defined as a *propositional attitude* – an agent's mental attitude to a proposition. For example, "the river is muddy" is a proposition (a declarative statement about the state of the world), whereas "I believe the river is muddy" is a mental attitude to that proposition. An understanding of mental states is likely to be highly useful in interpreting and predicting others' actions, and thus in responding appropriately. Difficulties with ToM have been suggested in a wide range of clinical conditions, including autism, schizophrenia, and anxiety disorders (Baron-Cohen, 1990; Baron-Cohen, Leslie, & Frith, 1985; Brüne, 2005; Frith & Corcoran, 1996; Washburn, Wilson, Roes, Rnic, & Harkness, 2016). If one wishes to understand this key transdiagnostic social symptom, a mechanistic understanding of the processes underlying ToM is crucial.

One prominent area of enquiry in seeking to understand ToM has been to examine the relationship between ToM and metacognition. Metacognition can be defined as 'cognition about cognition' (Georghiades, 2004) and, as such, can be considered as including meta-representations of one's own mental states – a form of self-directed ToM. Indeed, some researchers have considered ToM and metacognition as the same phenomenon (Gumley, 2011); in contrast, others have posited that metacognition and ToM are two distinct constructs which share a single cognitive system (Carruthers, 2009, 2011; Nicholson, Williams, Lind, Grainger, & Carruthers, 2020); whilst yet others suggest that the two abilities are completely distinct (Bang, Moran, Daw, & Fleming, 2022; Nichols & Stich, 2003; Proust, 2007).

There are three main schools of thought on the relationship between metacognition and ToM. One-system theories suggest a single metarepresentational system underlies both metacognition and ToM (Carruthers, 2009, 2011; Gopnik, 1993; Happé, 2003; Nicholson et al., 2020; Wilson, 2004). The two-system theory, in contrast, suggests that these distinct abilities are served by entirely distinct neural mechanisms (Nichols & Stich, 2003), meaning that it should be possible to find a double-dissociation between ToM and metacognitive abilities. The two-system account further suggests that the representation of one's own propositional attitudes ('I believe that…' / 'I think that…') is distinct both from the representation of one's own cognitive performance (such as in perception or memory tasks) and from the representation of others' propositional attitudes. A third theory states that metacognition is prior (Goldman, 2006), positing that the metacognitive system is recruited alongside other systems to infer the mental states of conspecifics. Specifically, the metacognition-is-prior theory suggests that to perform ToM, one must simulate oneself in the situation of the target and infer one's own mental state in those circumstances. As such, an inability to represent one's own mental state would severely impair both metacognition and ToM, whilst a ToM impairment would not be expected to impair metacognition.

Previous studies have addressed the relationship between ToM and metacognition and provided some, albeit mixed, evidence of a relationship between these two abilities. These studies typically measure participants' awareness of the accuracy of their responses in some first-order cognitive task (e.g., a perceptual or memory task) to assess metacognitive ability. Relative to much of the theoretical and philosophical work discussed above, this operationalisation used in experimental psychology is quite constrained, and it might be more precise to consider this work as seeking to examine the relationship between ToM and metacognitive sensitivity (the ability to discern the quality of one's cognitive performance). However, as we will discuss, the measurement of metacognitive sensitivity in much of this previous work is confounded with other variables. As such, throughout this paper, we will use the term 'metacognitive ability' for conceptual and general discussion, and the term 'metacognitive sensitivity' only when discussing the measurement of individuals' ability to discriminate accurate from inaccurate performance in a first-order task.

Many studies have found correlations between measures of metacognitive and ToM abilities (Carpenter, Williams, & Nicholson, 2019; Nicholson et al., 2020; van der Plas et al., 2021; D. M. Williams, Bergström, & Grainger, 2018), but this is not always the case (K. L. Carpenter et al., 2019). Similarly, whilst some studies have reported impairments in metacognitive ability associated with autism (Grainger, Williams, & Lind, 2016; Johnstone, Friston, Rees, & Lawson, 2022; Nicholson et al., 2020; van der Plas et al., 2021; Wilkinson, Best, Minshew, & Strauss, 2010; D. M. Williams et al., 2018; Wojcik, Moulin, & Souchay, 2013), a condition in which ToM is thought to be impaired (Abell, Happé, & Frith, 2000; Baron-Cohen, 1990; Baron-Cohen et al., 1985; Happé, 1994), other studies have failed to find such a deficit (K. L. Carpenter et al., 2019; Wojcik, Allen, Brown, & Souchay, 2011). Even amongst the studies in which an autistic metacognitive deficit has been observed, it has been seen in children but not adults (Wilkinson et al., 2010), in some tasks and not others (Wojcik et al., 2013), and when comparing diagnosed individuals with neurotypical adults but not when using continuous measures of autistic traits (D. M. Williams et al., 2018). Regardless, the results of studies suggesting deficits in both ToM and metacognition in autism have usually been interpreted as supporting the one-system view of metacognition and ToM, given that damage to a single system would lead to impairments in both abilities.

However, there are three possible explanations for data suggesting that ToM and metacognitive abilities are related, and that both are impaired in autism. First, it may be the case that ToM and metacognition are indeed subserved by a single system. In this case, the representation of propositional attitudes (mental states) would be a product of the same system as the representation of other forms of cognition, such as perception or memory.

Second, the apparent relationship between metacognitive and ToM abilities may be a product of some other factor which influences the measurement of both abilities in the relevant studies. As noted by van der Plas and colleagues (2021), many studies which have sought to test this relationship (e.g., (K. L. Carpenter et al., 2019; Grainger et al., 2016; D. M. Williams et al., 2018;

Wojcik et al., 2013)) make use of metacognitive measures in which metacognitive sensitivity (i.e., the extent to which confidence tracks accuracy) is not measured independently of metacognitive bias (i.e., the tendency, in general, to be more or less confident in responses), or perceptual or memory task performance. As such, the observed relationship between metacognition and ToM in these studies may be due to a third factor (such as confidence level or performance), leading to a spurious correlation between these abilities. This explanation appears all the more likely in light of evidence that autistic traits are associated with ToM ability (Abell et al., 2000; Baron-Cohen et al., 1985; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Dziobek et al., 2006; Happé, 1994), sensory sensitivity (relevant to perceptual task performance) (Ashwin, Ashwin, Rhydderch, Howells, & Baron-Cohen, 2009; Jussila et al., 2020; Takarae, Sablich, White, & Sweeney, 2016), and average confidence in task performance (McMahon, Henderson, Newell, Jaime, & Mundy, 2016; Z. J. Williams et al., 2022; Zalla, Miele, Leboyer, & Metcalfe, 2015).

To our knowledge, to date, there are only two studies which directly relate metacognition and ToM and have utilised measures of metacognitive sensitivity which are independent of metacognitive bias and cognitive performance. These studies are those by Nicholson and colleagues (2020), and by van der Plas and colleagues (2021). Although not the only way to dissociate metacognitive sensitivity from metacognitive bias and task performance, both studies measure metacognitive efficiency, which is defined as metacognitive sensitivity (measured in a bias-free manner) relative to first-order task performance (Fleming & Lau, 2014). The latter study claimed to have identified and resolved several potential problems with the former, including potential confounds of verbal fluency and response to ambiguous feedback. Van der Plas and colleagues provided evidence for a positive association between ToM ability and metacognitive efficiency, along with evidence that both ToM ability and autistic traits modulate the use of one's own behavioural cues (namely reaction time) in constructing confidence in one's own performance. These results are an important advance in explaining observed differences in metacognition across those with different levels of autistic traits or ToM ability, especially because they shed light on a possible mechanism through which these characteristics may relate to metacognitive ability (namely the use of visible cues in the construction of confidence).

However, the results of van der Plas and colleagues do not preclude the third possible explanation for the apparent relationship between metacognitive ability and ToM. It may be the case that metacognition is a useful tool in the complex process of making an accurate mental state inference (how ToM is tested), without the two abilities being served by a single system (as in the one-system view), and without metacognition being a *necessary* precursor to holding representations of the mental states of others (as in the metacognition-is-prior view). The notion that metacognition may neither make use of the same system as ToM, nor be a necessary precursor to ToM ability, but may still be useful in the process of ToM inference, may explain the mixed results observed in the literature (K. L. Carpenter et al., 2019; Grainger, Williams, & Lind, 2014; Grainger et al., 2016; Nicholson et al., 2020; van der Plas et al., 2021; Wilkinson et al., 2010; D. M. Williams et al., 2018; Wojcik et al., 2011; Wojcik et al., 2013).

A possible mechanism through which metacognitive ability may aid in ToM inference arises from consideration of the Mind-space framework (Conway, Catmur, & Bird, 2019). The Mind-space framework suggests that minds with different traits (relatively enduring individuating features such as personality traits or cognitive abilities) may give rise to different mental states in the same situation. This theory therefore predicts that traits should be a rich source of information when inferring an individual's mental state. Specifically, a mentaliser (a person making mental state inferences) may use a representation of a target's (the individual whose mental states are being inferred) traits to obtain an estimate of the target's mental state in a given situation (Conway et al., 2019). For example, if I believe that an individual is highly extraverted, I expect that at a party they will hope to speak to as many people as possible. Of course, a mental state (i.e., a propositional attitude held at a particular moment in time) need not always be wholly in line with one's typical responses (i.e., those that might be expected given one's traits) and can be influenced by situational factors. For example, an individual who typically wishes to interact with many others might actively seek interaction with only a specific individual at a particular party.

As such, the theory posits that, when making mental state inferences, mentalisers should make use of information about both the situation a target is in, and the traits of their mind. Trait inferences are thought to be represented through locating the target individual in Mind-space, a multi-dimensional space in which individual, non-orthogonal dimensions represent individual traits and their covariation. A target's location in this multi-dimensional Mind-space is therefore a mental representation of the qualities of the target's mind. The mentaliser can then combine their inferences about the target's mind with diagnostic situational information and reach a conclusion about their likely mental state, given the mentaliser's understanding of which mental states minds in that location give rise to in that situation.

Support for the Mind-space framework has come from experiments which demonstrate that manipulating participants' impressions of targets' traits, either directly or through manipulating impressions of related traits, affects participants' mental state inferences (Conway et al., 2020). Furthermore, it has been demonstrated that participants update their inferences about targets' mental states in line with updates to their perceptions of the targets' traits, in a manner that varies according to systematic relationships between traits and mental states (Long, Cuve, Conway, Catmur, & Bird, 2022). Importantly, the latter study demonstrated that the accuracy of specific mental state inferences is associated with the accuracy of specific trait inferences, again according to varied, but systematic, relationships.

Given evidence that people make use of inferences about target traits to inform inferences about target mental states, one might consider the role of metacognition in optimising the use of trait information. There is often some degree of error in trait judgments, and these errors may stem from different sources: one might have little or poor-quality information about a given trait, might be worse at inferring some traits than others, or might be more or less precise at different levels of traits (for example, trait inferences may improve when the target's location in Mind-space is close to the mentaliser's own (Conway et al., 2020)). If, as the evidence described above suggests, trait inferences are utilised to make mental state inferences, then erroneous trait inferences increase the risk of making erroneous mental state inferences, and thus misunderstanding others or behaving inappropriately. However, the converse is also true – making maximal use of highly accurate trait inferences facilitates more accurate mental state inferences.

A mentaliser's goal, then, should be to discount potentially misleading erroneous trait inferences and to maximise the use of helpful, accurate trait inferences. Metacognitive confidence is thought to facilitate the optimal use of information in the face of uncertainty (Fleming & Daw, 2017; Körding, 2007; Yeung & Summerfield, 2012), allowing us to place greater weight on higher confidence information and therefore, where confidence is positively related to accuracy, to rely more heavily on more accurate information. Under the Mind-space framework, this process should be particularly useful in minimising mental state inference error. If a mentaliser wishes to maximise the use of helpful trait information, they may rely more heavily on trait inferences in which they are more confident. In contrast, to minimise the introduction of error into mental state inference, they may down-weight trait inferences in which they are not confident.

Whether the process of weighting trait inferences according to confidence succeeds in improving the accuracy of mental state inference should therefore depend on the extent to which the mentaliser's confidence is a reliable indicator of the accuracy of their trait inference, i.e., their metacognitive sensitivity. Therefore, mentalisers with greater metacognitive sensitivity should generate more accurate mental state inferences than those with poorer metacognitive sensitivity. This hypothesis is illustrated in Figure 1. Following this line of reasoning, we postulated that the association between metacognition and mental state inference accuracy occurs because those who show higher metacognitive sensitivity are more able to adjust their use of trait inferences in line with the accuracy of those inferences, rather than (or in addition to) metacognition being necessary for holding representations of other's mental states, or these two abilities relying solely on a single system.

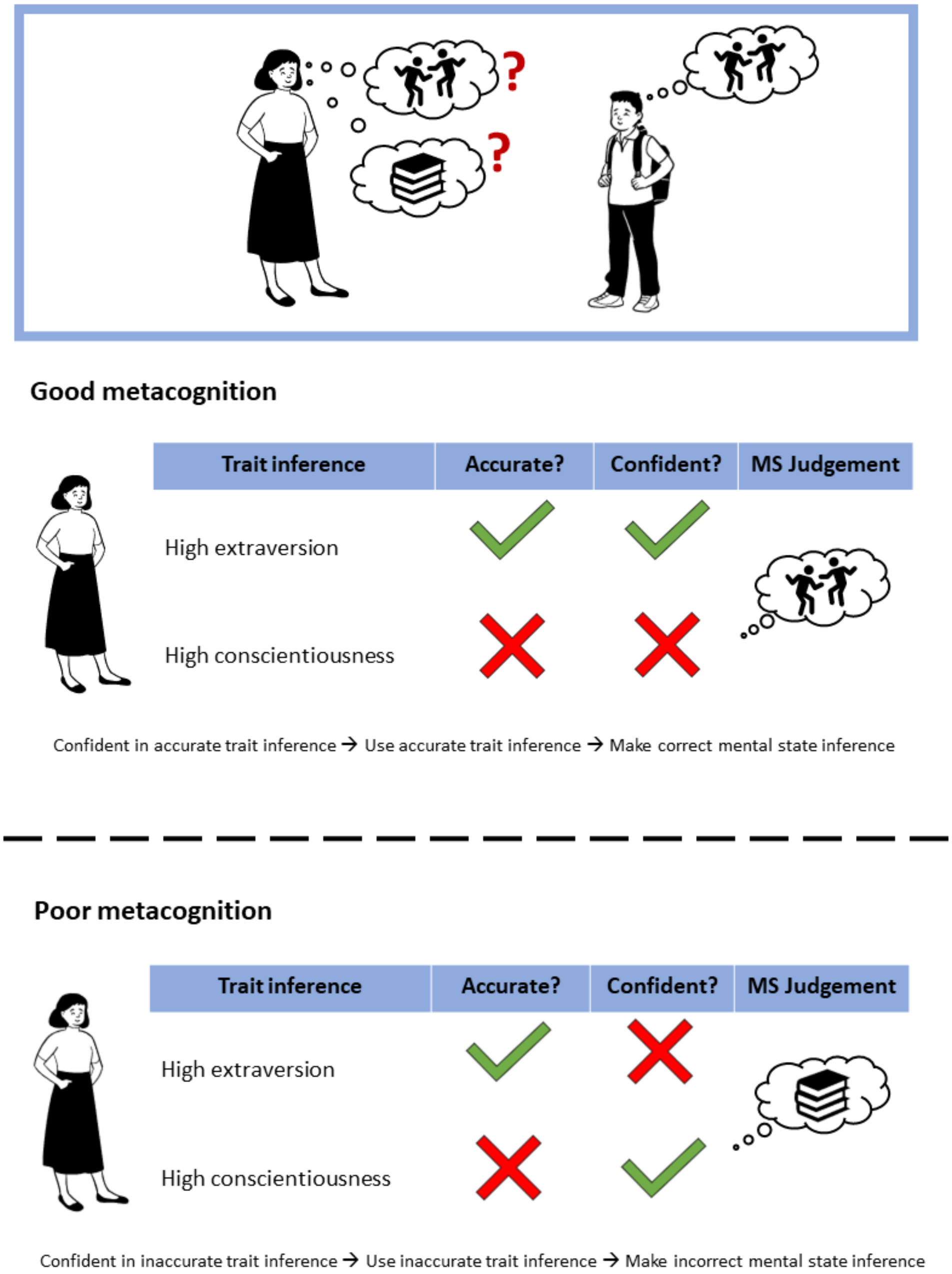

Figure 1. A schematic illustrating our hypothesis regarding metacognitive ability. Consider a teacher trying to infer her student's intention – to either go to a party or to do homework this weekend. This teacher believes (correctly) that a conscientious individual would intend to do the homework, and an extraverted individual would intend to go to a party. The teacher believes that the student is both highly extraverted and highly conscientious. The student is in fact highly extraverted and not

conscientious. If the teacher has high metacognitive ability, she will be confident in her accurate extraversion judgement and not in her erroneous conscientiousness judgement. She will base her inference on the accurate judgement and disregard the inaccurate judgement, to correctly infer that the student intends to go to the party. If the teacher has low metacognitive ability, she may be confident in her inaccurate conscientiousness judgement and not in her accurate extraversion judgement. She would then base her inference on the inaccurate judgement and disregard the accurate judgement, to incorrectly infer that the student intends to do the homework.

The present study seeks to test this theoretical explanation of the role of metacognition in mental state inference by examining the roles of both “first-order” trait inference ability and “second-order” metacognitive awareness of trait inference errors when deriving mental state inferences. Specifically, this study examines whether individuals weight their trait inferences according to their confidence, and, if so, whether this weighting process leads to differing levels of mental state inference accuracy in individuals with varying levels of metacognition. To do so, we made use of two tasks designed to resolve issues with commonly-used tasks.

First, we developed a novel metacognition task which tests metacognition specifically in the domain of trait inference. The question of the domain-generality of metacognition is still not resolved – there is evidence of behavioural dissociations in metacognitive abilities across domains in both healthy and clinical populations (Fitzgerald, Arvaneh, & Dockree, 2017; Fleming, Ryu, Golfinos, & Blackmon, 2014), suggesting some specificity; evidence that metacognitive training transfers across domains (J. Carpenter et al., 2019), suggesting some level of generality; and neural evidence suggesting both domain-general and domain-specific processes in metacognition (Morales, Lau, & Fleming, 2018; Rouault, McWilliams, Allen, & Fleming, 2018). It seems likely, then, that there may be some global metacognitive ability, but that domain-specific processes (which can be differentially effective) also exist.

As such, we ensured that metacognitive sensitivity was measured in the trait inference domain. The importance of doing so stems from the fact that our hypothesis regarding the role of metacognition in mental state inference refers specifically to the role of confidence in trait inferences, and the extent to which confidence in trait inference tracks the accuracy of those inferences. It is therefore crucial that domain-specific metacognitive sensitivity, above and beyond *general* metacognitive ability, is captured by our measure. In brief, our metacognition task utilised a judgement-of-confidence paradigm, in which participants rated their confidence in their trait inferences. Importantly, we also ensured that our measure of metacognitive sensitivity was independent of metacognitive bias (Fleming & Lau, 2014), resolving concerns regarding the role of average confidence levels (van der Plas et al., 2021).

Our second task was a recently-developed ToM measure, the Interview Task (Long et al., 2022). The Interview Task assesses the accuracy of mental state inferences against ground-truth information. Briefly, participants are presented with videos of unscripted practice job interviews and asked about the mental states of both targets (the interviewer and the candidate). For example, participants are asked ‘How would the candidate rate their performance in the interview?’ and ‘To what extent does the interviewer think that they put the candidate at ease?’. Participants’ judgements of the targets’ mental states are then compared to ground-truth information, obtained by having the targets report their mental states at the time of recording. Targets were not actors and were behaving freely within the context of the practice interview, meaning they were able to respond to one another however they wished and report their genuine mental states. As well as rating the targets’ mental states, participants were asked to rate the traits of the targets and their confidence in each of their trait judgements. Trait inference accuracy can then be assessed by comparing participant judgements to ground-truth information obtained through validated measures of the targets’ true traits.

The assessment of ToM ability through measuring the accuracy of inferences against ground-truth information is a substantial advantage of the Interview Task over other tasks in the ToM literature.

Typically, studies examining the relationship between metacognition and ToM (K. L. Carpenter et al., 2019; Nicholson et al., 2020; van der Plas et al., 2021; D. M. Williams et al., 2018) have made use of one or both of two tasks: the Reading the Mind in the Eyes Test (Baron-Cohen, Wheelwright, Hill, et al., 2001), and the Frith-Happé Animations Test (Abell et al., 2000). In both tasks, due to a lack of ground-truth information, the accuracy of participants' judgements, and thus their measured ToM ability, is determined by comparing their answers to 'correct' answers which are defined by the experimenter, or by the consensus of typical individuals. That is, participants are assessed against how other typical agents usually interpret the mental states, not against the mental states of the target agents themselves. As such, the Interview Task has the substantial benefit of having true correct answers derived from real agents, meaning both that ability is assessed in line with task instructions, and that the measured ability may be more likely to be reflective of true abilities outside of the laboratory setting. Furthermore, both the Reading the Mind in the Eyes Test and the multiple-choice version of the Frith-Happé Animations Test assess participants' inferences about agents' feelings and may therefore be truly assessing abilities other than ToM, defined as the inference and representation of propositional attitudes (Leslie & Frith, 1987; Oakley, Brewer, Bird, & Catmur, 2016).

Using the Interview Task, we can measure the accuracy and confidence of specific trait inferences about specific targets and examine the influence of those trait inferences on accompanying mental state inferences. By using the Interview Task alongside our novel metacognition task as well as a measure of autistic traits, we were able to test several hypotheses. First, we examined the association between autistic traits and metacognitive sensitivity. Given the equivocal nature of existing evidence surrounding this relationship (K. L. Carpenter et al., 2019; Grainger et al., 2016; Nicholson et al., 2020; van der Plas et al., 2021; Wilkinson et al., 2010; D. M. Williams et al., 2018; Wojcik et al., 2011; Wojcik et al., 2013), we did not have a specific prediction regarding this association. Second, we predicted that the previously observed association between trait inference accuracy and mental state inference accuracy would be replicated (Long et al., 2022). Third, according to the theory described above, we predicted that metacognitive sensitivity would predict mental state inference accuracy.

When examining the mechanism through which this association between metacognitive sensitivity and mental state inference accuracy may occur, we predicted that when participants reported higher confidence in a trait inference, any error in that trait inference would be more likely to be propagated into associated mental state inferences. As such, the relationship between trait inference error and mental state inference error should be stronger when confidence is high, as more of the error in trait inference is propagated to the mental state inferences than when confidence is low. We expected this functional relationship (a two-way interaction between trait inference error and confidence when predicting mental state inference error) to be present in those with both high and low metacognitive sensitivity. Statistically, however, we predicted the existence of a three-way interaction between trait inference error, confidence and metacognitive sensitivity when predicting mental state inference error, for the following reason.

With higher metacognitive sensitivity, indicating a better ability to discriminate between accurate and inaccurate trait inferences, high confidence trait inferences should be more accurate, and thus there should be less error to be propagated to the mental state inferences. Furthermore, error from low confidence trait inferences, which should be less accurate, will be less likely to be propagated; instead, the mental state inferences will be determined by other available information, including other more accurate trait inferences. As such, an individual with high metacognitive sensitivity should use trait inferences more optimally, such that mental state inferences are as accurate as possible given the available information. Statistically, if this is the case, then the close coupling of trait inference error and confidence should reduce the magnitude of the two-way interaction between trait inference error and confidence influencing mental state inference.

When metacognitive sensitivity is low, participants' confidence in their trait inference will be, by definition, less strongly related to the accuracy of that trait inference. With lower metacognitive sensitivity, then, trait inference error should be more evenly distributed across reported confidence levels, and the likelihood of that error being propagated to the mental state inferences should be determined by confidence. Therefore, we predicted that the modulatory effect of confidence on the relationship between trait inference error and mental state inference error would be larger in participants with lower metacognitive sensitivity. Specifically, the decoupling of confidence from trait inference error means that the role of confidence in determining the extent to which error is propagated should be more clearly observable in resultant mental state inferences, because the trait inference error that may or may not be propagated is more evenly distributed across levels of reported confidence, and it is therefore more likely that there will be error to propagate in high confidence trials.

The three-way interaction, then, is to be expected due to varying degrees of coupling between error and confidence as a function of metacognitive sensitivity but does not imply that there is a functional difference in the use of trait information and confidence in individuals with differing levels of metacognitive sensitivity.

The present study had the additional aim of examining factors which might be associated with the accuracy of trait judgements. First, we tested the association between participants' understanding of the traits of the 'average' mind (i.e., the median mind) and their trait inference accuracy. A positive association between error in the understanding of median traits and the error of trait inferences was expected for several reasons. Given that is it posited that both the structure of Mind-space and the ability to locate individuals within that space are experience-dependent (Conway et al., 2019), a better understanding of the population median might be reflective of experience interacting with a more representative group of individuals, which should aid the location of targets in Mind-space. Furthermore, an accurate understanding of the 'average' mind might reduce error by providing the most accurate possible 'default' inference when direct information about a given trait for a particular target is not available. Finally, an individual who tends to locate specific targets in Mind-space more accurately should be better able to intuit median population traits, on the basis that they have accurately located individuals they have encountered and can thus calculate the population median based on accurate data.

Additionally, we sought to further build upon a previous finding that a participant's trait inferences were observed to be more accurate when the target's traits were more similar to those of the participant (Conway et al., 2020). Conway and colleagues observed a similarity effect when participants saw thin-slice videos of targets of between six to nine seconds. We tested whether this effect would also be seen with longer videos, of approximately 30 seconds in the metacognition task and four minutes in the Interview Task. We could therefore establish whether the similarity effect is only present when there is very little information on which participants could base their judgement, or whether similarity continues to have a beneficial effect on trait inference accuracy when rich information about target traits is available. In line with previous predictions regarding the similarity effect (Conway et al., 2019; Conway et al., 2020), we expected that the effect would persist in longer videos, as the similarity effect is thought to reflect a greater ability to accurately locate individuals in Mind-space on the basis of behaviour when the targets' behaviour reflects one's own traits. Participants have a wealth of data about the behaviours associated with their own traits, due to the vast experience they have of themselves. As such, the similarity effect should occur regardless of the amount of information available in the stimuli, provided there is still some level of ambiguity and trait inference accuracy is not at ceiling.

However, the Mind-space theory suggests a possible limit to this similarity benefit which we sought to test in the present study. If the similarity effect can be explained by the wealth of information participants have about behaviours associated with their own traits, then target similarity should only be beneficial if participants can accurately represent their own traits (Conway et al., 2019). If not, participants may accurately recognise that targets are similar to them, but attribute to those targets the inaccurate traits they have attributed to themselves. These hypotheses are illustrated in

Figure 2. As such, whilst we expected to observe the similarity effect in trait inference across both tasks, we predicted that this effect would be modulated by the accuracy with which participants located themselves in Mind-space. Specifically, we predicted that the similarity effect would be stronger for those who were more accurate in their estimates of their own traits.

Having examined factors thought to influence the accuracy of trait inferences, we tested a final set of hypotheses examining whether these factors predicted not only the accuracy of trait inferences, but also metacognitive judgements about those inferences. If, as we suggest, mentalisers tend to make more accurate inferences about the traits of those who are more similar to them, we theorised that mentalisers might use similarity as a cue when determining their confidence in a trait inference. Given evidence that people tend to be overconfident in their own performance (Baranski & Petrusic, 1995; Brenner, Koehler, Liberman, & Tversky, 1996; Dunning, Griffin, Milojkovic, & Ross, 1990; Hoffrage, 2017; Moore & Schatz, 2017), we hypothesised that participants would be more likely to be overconfident when making a trait inference about a target who is more similar. As such, we predicted that participants would show a less negative relationship between error and confidence when making inferences about more similar targets.

Furthermore, if our hypothesis that the similarity effect is modulated by the accuracy of the mentaliser's perception of themselves is indeed correct, then use of this cue should be less effective for individuals with less accurate self-perception. In this case, a participant might be expected to accurately perceive a target to be similar to them and thus be more confident in their inference, but to mislocate the target in Mind-space due to their own erroneous self-perception. As similarity is less indicative of accuracy (and therefore a less useful cue for confidence) when self-perception error is higher, we hypothesised that any overconfidence effect (in which trait error is less negatively related to confidence when the target and participant are similar) would be larger when self-perception error is greater.

In sum, the present study sought to examine a possible mechanistic role for metacognition in ToM, testing the hypothesis that metacognitive abilities determine whether one can weight trait inferences optimally when deriving a mental state inference. The study also sought to test additional predictions about possible influences on the accuracy of trait inferences themselves. We suggested that individuals with more accurate understandings of the average mind would make more accurate trait inferences. We also predicted that participants would make more accurate trait inferences when the target is more similar to them, but that this similarity effect would be modulated by the accuracy of participants' understandings of their own traits. The final analysis of the study sought to examine whether these possible influences on the accuracy of trait inference affect confidence in trait inferences. We therefore tested the hypothesis that similarity is used as a cue in the construction of confidence, but that the degree to which this cue facilitates accurate metacognitive judgements is modulated by the accuracy of the individual's self-perception.

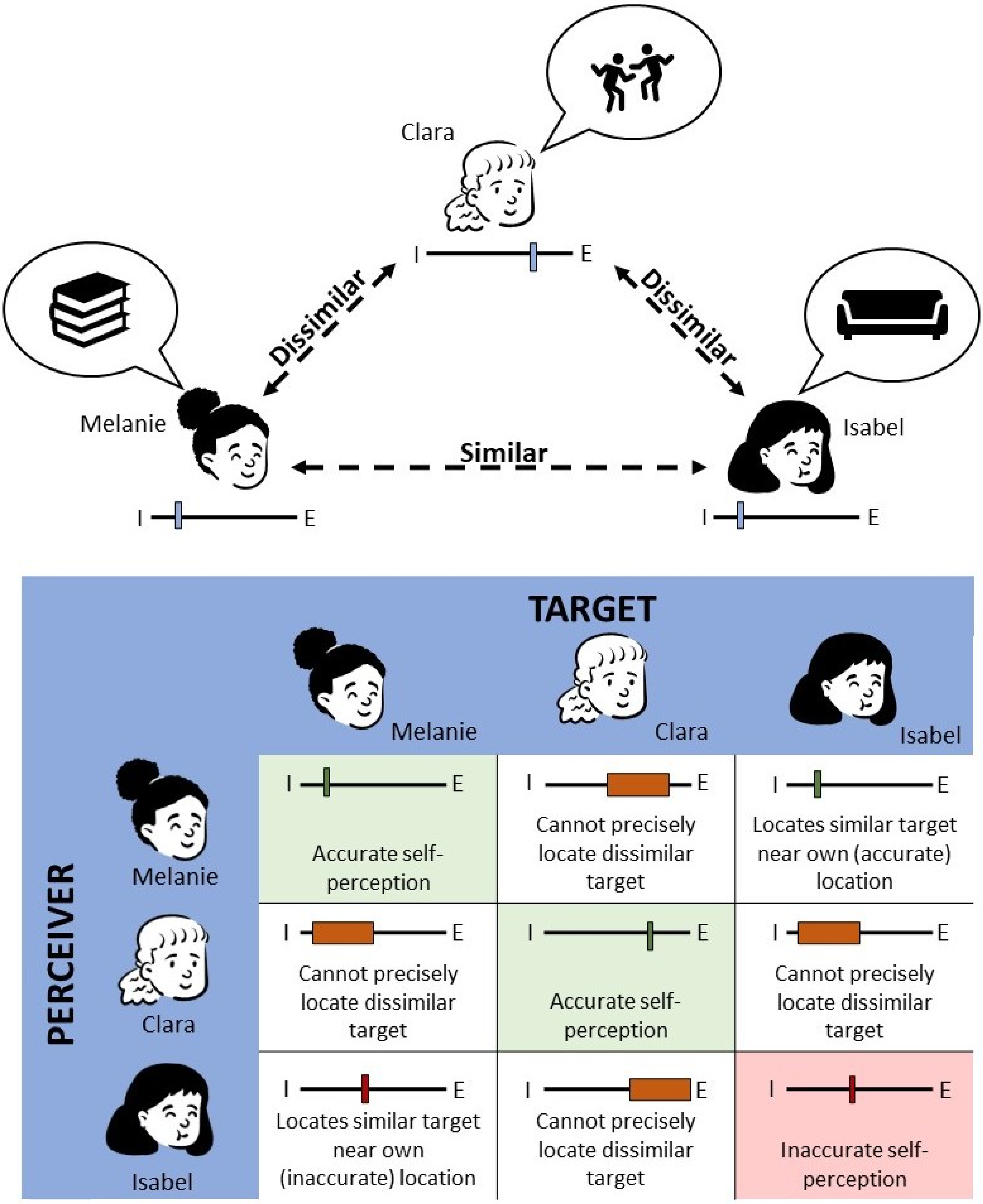

Figure 2. A schematic illustrating our hypotheses regarding similarity and self-perception accuracy effects on trait perception. Consider three classmates discussing what they did last weekend – the more extraverted classmate (Clara) went to a party, whilst one of the more introverted classmates (Melanie) did their homework and the other more introverted classmate (Isabel) watched TV. Each classmate's true level of extraversion is given on a scale (from I = introverted to E = extraverted)

beneath their picture in the triangle (top) and each classmate's judgements of themselves (shaded) and each other (unshaded) are given in the table (bottom). To illustrate our hypotheses, we will consider each perceiver in turn, following each row of the table to understand their judgements of themselves and others. Melanie has accurate self-perception. She recognises that if she had not had homework to do, she would have behaved like Isabel, so accurately locates Isabel near herself in Mind-space. She may infer that Clara is more extraverted than her but has less information about what Clara's behaviour suggests of her precise level of extraversion. Clara has accurate self-perception. However, she would not behave like either Melanie or Isabel. She therefore has little information available to allow her to interpret their behaviour in terms of their introversion. Isabel has inaccurate self-perception. She recognises that if she had had homework to do, she would have behaved like Melanie. She locates Melanie near herself in Mind-space but, because she believes herself to be more extraverted than she truly is, overestimates Melanie's extraversion in accordance with her erroneous self-perception. She would not behave like Clara so, again, has little information on which to base a precise inference.

2. Methods

2.1. Participants

92 participants completed the experiment. Volunteers participated online through the website prolific.co and were compensated for their time. Four participants were excluded as their responses suggested that they failed to engage with the task. Specifically, these participants gave identical confidence ratings for over 90% of trials on one or both of the primary tasks (the metacognition task, and the Interview Task). Five participants scored zero or one out of four on basic factual questions in the Interview Task. These questions were designed as attention checks rather than control questions and as such these participants were excluded. Five further participants were removed in the process of outlier exclusion (see Section 2.3.2 below).

The remaining 79 participants (46 female) had a median age of 27 years (SD = 8.82), and all participants were over 18 years old. All participants gave informed consent online and the study was approved by the University of Oxford Central University Research Ethics Committee and followed the principles of the Declaration of Helsinki. One of these participants did not provide responses to the personality questionnaire, and so was excluded only from analyses requiring the missing data – i.e., analyses examining the predictors of trait inference accuracy.

2.2. Procedure

The experiment was hosted on gorilla.sc (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). Each participant completed all components of the experiment across two sessions with a one-day delay between sessions. Each session took approximately one hour. It was not crucial that participants had a standard delay between sessions, as it was not thought that the delay would affect performance on the second task (for example, there was no memory component). Instead, the delay served to enforce a significant rest-break for participants to avoid fatigue in the second part of the experiment.

In the first session the participants completed the metacognition task and associated post-task questions. In the second session participants completed the Autism-Spectrum Quotient measure AQ-Short (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001; Hoekstra et al., 2011), 20-item Toronto Alexithymia Scale (Taylor, Bagby, & Parker, 2003), the HEXACO-60 PI-R personality inventory (Ashton & Lee, 2009), and the matrix reasoning portion of the Wechsler Abbreviated Scale of Intelligence (Wechsler, 2011) as well as the Interview Task, which measures mental state inference accuracy.

2.2.1. Metacognition task

Participants were first presented with instructions for the metacognition task and asked three multiple-choice questions about those instructions. If participants answered any of these questions incorrectly, they would be presented with the instructions again and asked the same questions in a new randomised order. Participants had three attempts to answer the instructions quiz correctly – participants who failed on the third attempt were not allowed to continue the experiment.

On entering the task, participants were presented with descriptions of the six HEXACO personality dimensions (honesty-humility, emotionality, extraversion, conscientiousness, and openness-to-experience (Ashton & Lee, 2007; Ashton, Lee, & De Vries, 2014; Lee & Ashton, 2008)), taken from https://hexaco.org/scaledescriptions. In the task itself, participants watched 21 videos in which an interview candidate answered the prompt 'Tell me a little about yourself.'. These videos were shortened clips from the video stimuli used in the Interview Task, and each was 20-30 seconds long. When editing, it was ensured that videos stopped the first time the candidate reached the end of a sentence once the initial 20 seconds had passed. These videos were arranged into seven blocks of three videos and both block order and trial order within block were randomised.

On each trial, participants watched the video and were asked to rate the candidate on the HEXACO personality dimensions. These ratings were given along a continuous slider on which the left-hand side represented a score of zero and the right-hand side a score of 100. The zero to 100 scale was used to allow participants to give precise scores which were later converted to the one to five scale used in the HEXACO-60 (Ashton & Lee, 2009). The start-point was the midpoint of the slider and participants did not see the numerical score they were giving to the target. Participants also gave a judgement of their confidence in each personality rating they made on a one to five scale, with one indicating low confidence and five indicating high confidence. Judgements could be made whilst the video was playing and could be adjusted until the participant chose to progress the screen. Participants could not progress until they had viewed the entirety of the video but had unlimited time to respond once the video had ended. The videos could not be replayed.

Each block also contained an attention check trial at a random point within the block. On these trials, no video played and the text of the questions, which usually read, for e.g., 'How conscientious is the candidate?' and 'How confident are you in your answer?', instead said 'Move the slider all the way to the right.' and, for e.g., 'Press 3'. The number which participants were instructed to press for their confidence rating varied for each attention check trial. Participants who selected an incorrect number on two or more attention check trials during the task were not allowed to continue the experiment.

Following the task itself, participants were asked several questions. First, they were asked to rate 'the average person' on each of the six HEXACO dimensions. Then, they rated themselves on the same dimensions. These measures were later used to establish participants' accuracy in their perception of the median person's traits, and of their own. Participants were also asked about the standards against which they assessed target traits. Given that there was little variation in reported strategy (54 participants reported comparing the target to the average person, 17 reported comparing the target to themselves, and seven reported another strategy), any analyses of these data would be underpowered and so these data will not be discussed further.

2.2.2. Interview Task

The Interview Task was designed to measure mental state inference accuracy (and thus ToM ability) against ground-truth information (Long et al., 2022). Participants viewed video stimuli of targets engaging in a practice job interview. The videos, each of which is between two and six minutes long, show an online video interaction between two individuals who were assigned to be the interviewer or candidate. These individuals were not actors and the interaction itself was not scripted. Interviewers asked three general set questions of the kind used in job interviews and were invited to ask any follow up questions they wished of the candidates. Each participant saw four videos which were randomly selected from a pool of twelve. On some trials, participants may have seen targets that they had seen previously in the metacognition task. However, memory effects were unlikely given that these targets were only seen for 20-30 seconds within a set of 21 clips at least one day prior to completion of the Interview Task. After each video, participants were asked a multiple-choice factual question about the content of the conversation, such as “What activities does the candidate say she likes?”. Participants who failed the factual question on three or more trials were excluded from the analysis. These questions were designed to assess whether participants were attending to the content of the videos; they were therefore very simple and were used only to exclude participants thought not to be attending to the stimuli.

At the end of each video, participants were asked to rate the candidate and interviewer on the HEXACO six personality dimensions, using the same slider system as in the metacognition task. They were then asked a series of questions about the mental states of the interviewer and the candidate. Participants answered 48 mental state questions in total, split evenly between questions about the interviewer and the candidate. These questions were answered in a continuous manner along sliders and are given in the Supplementary Materials (Section S.1.). The sliders had a scale of zero to 100 and the start-point was the centre of the slider.

Importantly, all candidates and interviewers completed the HEXACO-60 personality questionnaire (Ashton & Lee, 2009) and reported their mental states during the interview (for full details of the stimulus development procedure, see Long et al. (2022)). This means that the accuracy of participants’ personality and mental state inferences could be assessed against ground-truth data. Targets were asked to report their mental states on the same quantitative scale that participants later used to infer them, meaning that discrepancies between target and participant-inferred mental states were not binary, but continuous. Mental state inference error was obtained by taking the absolute difference between the ground-truth rating given by the target of the inference and the inferred rating given by the participant. Trait inference error for each trait was calculated as the absolute difference between the ground-truth value obtained from the target’s HEXACO-60 responses and the participant’s rating of the target’s trait.

2.2.3. Additional measures

Participants completed the AQ-Short (Hoekstra et al., 2011), the TAS-20 (Taylor et al., 2003), the HEXACO-60 PI-R (Ashton & Lee, 2009), and the matrix reasoning portion of the WASI-II (Wechsler, 2011). The AQ-Short is a 28-item version of the Autism-Spectrum Quotient (Baron-Cohen, Wheelwright, Skinner, et al., 2001). Participants rate the degree to which they agree they experience certain autistic traits. For each question, responses are on a scale between one (definitely disagree) and four (definitely agree). AQ scores are obtained by reverse scoring the necessary items and then summing the item scores to give an AQ score between 28 (minimum) and 112 (maximum).

The TAS-20 is a measure of alexithymic traits. Alexithymia is a sub-clinical condition in which individuals have difficulties interpreting their own emotions (Sifneos, 1973) and which often co-occurs with autism (Hill, Berthoz, & Frith, 2004). The TAS-20 was included in the current study as evidence suggests that emotional symptoms conventionally attributed to autism can actually be better explained by comorbid alexithymia (Bird & Cook, 2013). It was not expected that there would be an association between alexithymic traits and metacognition, but we chose to include alexithymia as a covariate to ensure that any observed differences are attributable to autistic traits themselves. On the TAS-20, participants rate the degree to which they agree that they experience various alexithymic traits. Responses on each question range from one (completely disagree) to five (completely agree). Again, a score is obtained by reverse-scoring necessary items and then summing the item scores to give a TAS score between 20 (minimum) and 100 (maximum).

The HEXACO-60 is a 60-item version of the HEXACO PI-R (Lee & Ashton, 2004). It measures personality along the six HEXACO personality dimensions: honesty-humility, emotionality, extraversion, conscientiousness, and openness-to-experience. Participants rate the degree to which they agree with statements about their behaviours and responses to certain situations on a scale between one (strongly disagree) and five (strongly agree). Factor scale scores are obtained by reverse scoring necessary items and then taking the mean across all ten questions loading onto that factor. This gives a score on each dimension between one (minimum) and five (maximum).

The matrix reasoning portion of the WASI-II was used to estimate intelligence. Intelligence was included as a control variable to ensure that any observed effects were not dependent upon any relationship between intelligence and autistic traits, alexithymic traits, or metacognition. The matrix reasoning portion of the WASI-II involves seeing matrices of images and choosing an image that fits a blank space in the matrix based on the rules governing the images and their placements. Participants were given two practice rounds in which they were given feedback. Participants then completed up to 30 trials with no feedback, but the task ended as soon as they had responded incorrectly to three consecutive trials.

2.3. Analysis Strategy

2.3.1. Statistical power

An a priori power analysis was conducted to determine the minimum sample size required to achieve 80% power when testing for an association between metacognitive sensitivity and mental state inference accuracy. The power analysis was conducted using G*Power 3.1.9.7. (Faul, Erdfelder, Lang, & Buchner, 2007). This indicated that using a one-tailed test with a medium effect size of $r = -.30$ (Cohen, 1988, 1992) and a significance criterion of $\alpha = .05$, the minimum sample size required for 80% power is $N = 64$. The obtained sample size of $N = 79$ is therefore adequate to test for the presence of this effect.

2.3.2. Outlier detection and removal

In the metacognition task, we excluded outlying datapoints which indicated that participants were not correctly engaging with the task or not paying sufficient attention to stimuli. As such, we excluded outlying observations of metacognitive sensitivity below the lower quartile, and outlying observations of mean trait inference error (across all targets) above the upper quartile. Outliers were defined as 1.5 times the interquartile range above the upper quartile or below the lower quartile. One participant was excluded as an outlying observation of metacognitive sensitivity, and four further participants were excluded as outliers in mean trait inference error.

The outlying AUROC2 score below the lower quartile was below 0.5 (AUROC2 = 0.43), and thus below chance, meaning that the participant in question consistently gave higher confidence ratings for inaccurate trait judgements, and lower confidence ratings for accurate trait judgements (for details see Section 2.3.3. below). There is no clear basis on which one would expect a participant to behave in this way and, as such, it is likely that this value is indicative of response error. As there was no equivalent reason to suspect that AUROC2 outliers above the upper quartile were non-legitimate, these observations (two participants) were retained. All outlying observations of trait inference error in the metacognition task represented high degrees of error and thus indicated possible inattention to the stimuli.

In the Interview Task, participants who had passed the factual attention check questions should be assumed to have attended to the task, and there was no basis on which to believe that outlying observations in this task were illegitimate (in contrast to the metacognition task, in which an outlying measurement indicated systematic mischaracterisation of accurate and inaccurate trials). As such, participants were not excluded on the basis of outlying performance. However, for the purposes of our mechanistic analysis, in which data from individual trials were analysed, observations were excluded on the trial level. Outlying observations of both mental state inference error and trait inference error were excluded. Outliers were again defined as observations lying more than 1.5 times the interquartile range above the upper quartile or below the lower quartile. No participant had more than 10% of their mental state or trait judgements judged as outliers, and so no participants were excluded on this basis. To ensure consistency, outlying observations were not included when calculating participant mean mental state inference error or trait inference error.

2.3.3. Metacognitive sensitivity analysis

Because little empirical work has used personality inference as a first-order task in the study of metacognition, we avoided using parametric measures such as meta-d’ or M-ratio as we could not be certain that necessary assumptions could be met. Specifically, the gold-standard metacognitive measure, meta-d’, and the accompanying M-ratio measure, relies upon an equal-variance Gaussian assumption for the underlying ‘type 1’ distributions of internal signal strength. The complexity of the stimuli and cognitive processes involved in trait inference, which is known to vary in difficulty according to characteristics of both the target and the participant observer (Conway et al., 2020), means that a non-parametric approach is most appropriate for assessing metacognitive sensitivity in the trait inference domain. Similarly, a two-alternative forced choice task is a requirement for fitting the signal detection theory model that underpins meta-d’ analysis, but such an approach is inappropriate in the trait inference domain, in part due to complexities in defining relative difficulty of trait inference. As such, we used the area under the type 2 receiver operating characteristics curve (AUROC2) method recommended by Fleming and Lau (2014) for cases where non-parametric analysis is most appropriate. AUROC2 is a bias-free metric of the extent to which confidence distinguishes between correct and incorrect trials (Clarke, Birdsall, & Tanner Jr, 1959).

In order to obtain binary trait inference performance, responses were converted from a continuous scale to a binary metric which indexed whether the participant placed each target above or below the population median on the specific personality dimension in question (using data obtained by Ashton and Lee (2009) for the population medians). Participant ratings of target traits were scored as either correct or incorrect based on whether they had rated the target as above or below the median and the true location of the target relative to the median on that personality dimension. The type 2 ROC curve was constructed for each participant by setting varying thresholds for categorising a response as ‘confident’ based on the confidence rating given by the participant. Specifically, the thresholds used for constructing the type 2 ROC curves were such that the first point took a confidence rating of 1 as ‘low confidence’ and anything higher as ‘high confidence’, the second took a confidence rating of 1 or 2 as ‘low confidence’ and anything higher as ‘high confidence’ and so on and so forth. At each possible threshold, the participant’s type 2 hit rate (i.e., the probability of responding ‘confident’ given the trait judgement is correct) was plotted against the participant’s type 2 false alarm rate (i.e., the probability of responding ‘confident’ given the trait judgement is incorrect). The area under the resulting curve (the AUROC2) was calculated to give a measure of metacognitive performance.

An AUROC2 value of 0.5 indicates that the participant is as likely to make a type 2 false alarm judgement as a type 2 hit judgement, meaning that their metacognitive performance is at chance. To check that participants were performing above chance on the metacognition task, we assessed whether the mean value of the AUROC2s was greater than 0.5. To do this, we computed a one-sample t-test. The null hypothesis was that the population mean is 0.5. This test was one-tailed, testing the alternative hypothesis that the mean was greater than 0.5, as there is no reason to believe that participants would systematically misclassify performance in the manner required for the AUROC2 to be below 0.5.

2.3.4. General approaches for statistical modelling

To test our hypotheses, we conducted several statistical analyses, detailed below. In all cases, descriptive statistics indicated acceptable skew and kurtosis. All predictor variables were standardised by subtracting the sample mean and dividing by the standard deviation to aid interpretability.

Several of our analyses involved fitting linear mixed effects models using the lme4 (Bates, Mächler, Bolker, & Walker, 2014) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017) packages in R (R Core Team, 2020). In each case, we report the dependent variable, structure of random effects, and fixed effect predictors. For all linear mixed effects analyses, a model comparison approach was adopted. Broadly (and unless otherwise specified), this approach involved first fitting a null model which included only the random intercepts as predictors (null); then a model including our predictor(s) of interest, but with only random intercepts (intercepts-only); and finally, a model including any random slopes which are justified both by the experimental design and by the data (random slopes).

The structure of the random slopes model was determined by first fitting the maximal model. The maximal model was constructed according to principles outlined by Barr et al. (2013) – namely, slopes were included where doing so would not make the model unidentifiable. Specifically, slopes for variables that were obtained on a by-participant basis (AQ, TAS, WASI-MR, AUROC2 and mean trait accuracy in metacognition task) were not allowed to vary by participant. Once the maximal model had been fitted, the model was simplified according to the variance explained by each slope using the rePCA function in lme4, following Bates et al. (2015). This approach was used provided convergence was achieved. We report any convergence issues below.

Model comparisons were then performed using the Akaike Information Criterion (AIC), where a -2 difference indicates a significantly better fit (Burnham & Anderson, 2004). The AIC is a comparison method which penalises complexity and so was used to prevent overfitting. Given the large number of levels in our random effects (specifically, the fact that participants answered questions about 48 distinct mental states and 12 distinct traits for each of four videos), the Bayesian Information Criterion (BIC) was thought to be too conservative as a method of comparison (Dziak, Coffman, Lanza, Li, & Jermiin, 2020). All models compared, and their accompanying comparison statistics, are reported in the Supplementary Materials (Sections S.2. – S.5.).

Once the best fitting model had been determined, we used the summary function of lmerTest (Kuznetsova et al., 2017) to obtain coefficients and perform t-tests using Satterthwaites' method for degrees-of-freedom. We also report 95% confidence intervals calculated by bootstrapping with 500 simulations. To avoid issues with multi-collinearity and facilitate interpretability, mixed-effects models were fitted with x-standardisation (i.e., the predictor variables, but not the dependent variable, were standardised) and estimates are thus given in terms of the units of the dependent variable. For example, estimates arising from models of confidence are expressed as the predicted change in confidence rating (on the original 1-5 scale) for each standard deviation change in the predictor variable. We denote estimates of this kind as *B.* In contrast, to facilitate comparison with other work in the literature, beta coefficients obtained through linear regression are given in their full standardised form (denoted by *β*). Thus, these estimates express the predicted change in the dependent variable (in standard deviations) arising from each standard deviation change in the predictor variable.

2.3.5. Theory of Mind and metacognition analyses

First, we sought to establish whether there was an association between participants' measured metacognitive sensitivities (i.e., their AUROC2 scores) and our measures of autistic traits, alexithymic traits and intelligence. To do so, we conducted a linear multiple regression with participant AUROC2 scores as the dependent variable and AQ, TAS, and WASI-MR scores, as well as participant mean trait inference error in the metacognition task, as predictors. Theoretically, the AUROC2 measure of metacognitive sensitivity is not performance-independent, such that we should expect people who perform better on the trait inference task to show higher AUROC2 values even if they do not differ in metacognitive capacity (Clarke et al., 1959; Galvin, Podd, Drga, & Whitmore, 2003). To control for such dependence, participant mean trait inference error in the metacognition task was included in the model and a one-tailed test of its significance is reported. The same control was included in all analyses including metacognitive sensitivity.

Next, we examined predictors of ToM performance by conducting a linear multiple regression with participant mean mental state inference error as the dependent variable and the participant mean trait inference error in the Interview Task, AUROC2 score, and mean trait inference error in the metacognition task as predictors. Given existing evidence leads to the directional predictions that trait inference error should be positively related to mental state inference error and metacognitive sensitivity should be negatively related to mental state inference error, we report one-tailed tests for these variables.

We then tested whether the data supported our hypothesis regarding the mechanism of any relationship between metacognitive sensitivity and mental state inference error. We predicted that in participants with poor metacognitive sensitivity, confidence would modulate the effect of trait inference error on mental state inference error, such that error in trait inference should be more positively associated with mental state inference error when confidence is high than when it is low. This two-way interaction effect of confidence and trait inference error on mental state inference error was expected to be reduced in participants with good metacognitive sensitivity, resulting in a predicted three-way interaction effect including metacognitive sensitivity.

For this analysis, linear mixed effect models were fitted using the lme4 (Bates et al., 2014) and lmerTest (Kuznetsova et al., 2017) packages in R (R Core Team, 2020). Absolute mental state inference error was the dependent variable and random intercepts for participant, video, and trait-mental state combination (hereafter, trial) were included. First, we fitted a null model including the random intercepts as the only predictors. Then, we fitted a series of nested models including trait inference error on the given trial of the Interview Task, participant AUROC2 score, and participant mean trait inference error in the metacognition task. This allowed us to confirm that previously observed effects were also present when analysing the data in a trial-by-trial manner and to test whether our model of interest outperformed models including these main effects. Finally, we fitted the model of interest, an intercepts-only model in which the predictors were: trait inference error on the given trial of the Interview Task, associated reported confidence for that trial, participant metacognitive sensitivity (AUROC2), and participant mean trait inference error on the metacognition task. All variables except for participant mean trait inference error on the metacognition task were allowed to interact and the three-way interaction between metacognitive sensitivity, confidence and trait inference error was the primary term of interest. Given the complexity of this model, there was no principled way to determine random slope structure, and so an intercepts-only model was deemed most appropriate. Aside from the lack of a random slopes model, this analysis followed our model comparisons procedure outlined in Section 2.3.4. above.

Next, we sought to examine the relationships between performance on our two tasks. It is possible that the processes underlying trait inference and confidence in those inferences could differ when there is little information available (i.e., in our shorter videos in the metacognition task) compared to when there is more information on which to base inferences (i.e., in the longer Interview Task videos). Therefore, we conducted tests to establish whether individual differences in our metacognition task were associated with individual differences in processes underlying judgements in the Interview Task. To do so, we conducted three additional analyses. In the first, we examined the process of trait inference by testing the Pearson's product-moment correlation between mean trait inference error in the metacognition task, and mean trait inference error in the Interview Task.

The second analysis determined whether metacognitive performance in the metacognition task was associated with metacognitive performance in the Interview Task. We computed the Pearson's correlation between trait inference error and confidence separately for each task for each participant as a measure of metacognitive ability. This correlation measure is more likely to be confounded with metacognitive bias than the AUROC2 measure but provided comparable proxy measures of metacognition across both tasks. We then extracted the two Pearson correlation coefficients for each participant and tested the Pearson's correlation between the two correlations.

Finally, we compared participants' trait inferences in the Interview Task with their inferences about the traits of the same target in the metacognition task. To do so, we extracted trials of the metacognition task and the Interview Task in which participants assessed the same targets. Then we computed linear mixed effects models with Interview Task trait judgement as the dependent variable, and random intercepts for participant, video, and trait. We fitted a null model including only the random intercepts and an intercepts-only model including the participants' judgement of each trait for each target in the metacognition task as a predictor. We also carried out the procedure detailed above for determining the maximal models that are justified by the data. Whilst it was trial-by-trial trait inference errors, not judgements, that were used in our main analyses, the relationship between participants' judgements of each target gives the lower bound for the effect size of the possible relationship between participants' errors in those judgements. Specifically, if a participant underestimated a given trait for a given target in the metacognition task but overestimated it in the Interview Task, the errors may still be correlated if the absolute magnitude of the error is consistent. Cross-task analysis of trial-by-trial judgements is thus a more conservative measure of whether the trait inference process differed across the two tasks.

2.3.6. Trait inference accuracy analyses

Our second set of analyses addressed our hypotheses regarding factors associated with trait inference accuracy. We obtained a measure of the accuracy of participants' understanding of the average mind by computing the absolute difference between their rating of the average person for each trait and the population median value of that trait (using data obtained by Ashton and Lee (2009) from a Canadian student sample). We also obtained a measure of the accuracy of each participant's self-perception by computing the absolute difference between their rating of themselves on each trait dimension and their ground-truth score for that trait, obtained through scoring their responses to the HEXACO-60 questionnaire (Ashton & Lee, 2009). Participant average error in median rating and participant average error in self-perception were obtained by taking the mean of the participant's error in each domain across all traits.

To determine the relationship between the accuracy of a participant's understanding of the average mind and their performance on our tasks, we performed three multiple linear regressions. Each regression had a different dependent variable, reflecting the different types of performance which may be associated with understanding of the average mind. The first model included participant mean error in median rating and participant mean error in self-perception as predictors of participant mean trait inference error in the metacognition task. The second model included the same variables as predictors of participant mean trait inference error in the Interview task, and the final model included the same variables again as predictors of participant mean mental state inference error in the Interview Task. To assess whether any effect on mental state inference error was due solely to effects on trait inference error, we fitted a model in which the dependent variable was mean mental state inference error in the Interview Task and mean trait inference error in the Interview Task was included as a predictor in addition to participant mean error in median rating.

Next, we tested our hypothesis that the previously established similarity effect (Conway et al., 2020), in which participants make more accurate inferences about the traits of similar targets compared to dissimilar ones, would be modulated by the accuracy of participants' self-perception. Once again, linear mixed effect models were fitted for this analysis. The same process was carried out to test this hypothesis in both our metacognition task and the Interview Task. In both cases, the absolute error in participant trait inference for a given trait and target was the dependent variable and random intercepts were fitted for participant, video, and trait. Three models were fitted for model comparison, following the same process previously described. First, we fitted the null model, including the random intercepts but none of our predictors of interest. Then, as our measure of similarity, we included the absolute difference between the target's HEXACO-60 score and the participant's HEXACO-60 score for the same trait. Next, we added the absolute difference between the participant's rating of themselves and their HEXACO-60 score for the same trait, as our measure of self-perception error. Participant-target difference and self-perception error were allowed to interact in this model. Finally, we computed a random slopes model by completing the previously outlined process.

2.3.7. Predictors of confidence in trait inference

Our final analysis tested our hypothesis regarding how factors associated with trait inference accuracy might be related to participants' confidence in their trait inferences. For this analysis, we made use of the measures of participant self-perception error (the absolute difference between the participant's rating of themselves and their HEXACO-60 score for the same trait) and participant-target dissimilarity (the absolute difference between the target's HEXACO-60 score and the participant's HEXACO-60 score for the same trait) calculated for the mixed model analysis of metacognition task data outlined in Section 2.3.6. Linear mixed effects models were fitted with participants' reported confidence in their trait judgements in the metacognition task as the dependent variable, and random intercepts for participant, video, and trait.

Here, we followed the same procedure as used in the mechanistic analysis outlined in Section 2.3.5. First, we fitted a null model, including only the random intercepts as predictors. Following this, we fitted a series of nested models including trait inference error on the given trial of the metacognition task, participant-target trait difference, and participant self-perception error. In the model of interest, all three of these predictors were included and allowed to interact, and the interactions were the terms of interest. We predicted that participant-target difference and self-perception error would interact with trait inference error to determine confidence, reflecting an influence of these two predictors on metacognitive ability. Once again, given the complexity of this model and the predicted effects, there was no principled way to determine random slope structure, and so an intercepts-only model was deemed most appropriate.

As confidence ratings, the dependent variable in these analyses, took the form of a one to five integer scale, we conducted these analyses using two approaches. For our primary analysis, we treated confidence as a linear continuous variable, fitting our mixed-effects models using the lme4 package in R (Bates et al., 2014). However, we also conducted a supplementary analysis in which we treated confidence as an ordinal variable. For this, we fitted our models using the clmm function in the ordinal R package (Christensen, 2023). This approach involved fitting cumulative link mixed models, using a logit link function and allowing the threshold for each response category to vary. Models fitted using this approach predict the probability of each response (1,2,3,4, or 5) being given, without assuming that the thresholds for giving one response rather than the next are evenly spaced. This supplementary analysis was conducted to account for the integer nature of the confidence ratings, and to allow for the possibility that thresholds might differ between confidence levels (e.g., participants might require a greater increase in confidence to respond ‘5’ instead of ‘4’, than to respond ‘2’ instead of ‘1’). Both methods gave the same inferential results and, as there is no reason to suspect that participants did use differing thresholds (or that, if they did, those thresholds would be uniform across participants), the linear approach is reported here. The results of the ordinal analysis are given in the Supplementary Materials (Section S.5.).

3. Results

Descriptive statistics for metacognitive sensitivity (AUROC2) scores, mean trait inference error in the metacognition task, and covariates are given in Table 1. As shown in Figure 3, the type 2 ROC curves for most participants bow to the top-left of the diagonal line which represents chance performance, corresponding to an AUROC2 of 0.5. A one-tailed, one-sample t-test showed that the mean AUROC2 was significantly greater than chance, $M = 0.56$, t (78) = 10.79, $p < .001$, indicating that, on average, participants had significant insight into the accuracy of their trait inference judgments.

As previously mentioned, the participant with an AUROC2 score identified as an outlier below the lower quartile was excluded. Figure 3 shows that a small number of participants had AUROC2 scores which were below 0.5, but which were not outliers. If a participants’ confidence ratings were truly random (i.e., if they were equally likely to respond with high or low confidence regardless of the accuracy of the judgement), there would be a 50% probability of obtaining an AUROC2 value below 0.5, with the probability of obtaining a given value decreasing as that value deviates further from 0.5. As such, whilst these participants may have had little to no metacognitive insight into the accuracy of their trait inference judgments, these AUROC2 values are sufficiently close to 0.5 that it cannot be claimed with confidence that their scores are a result of response error. In addition, these participants passed the embedded attention checks in the Interview Task and were not classified as outliers for poor performance on the trait inference element of the metacognition task. These participants were therefore thought to be paying sufficient attention to the task and their AUROC2 score was thought to be reflective of their ability, albeit with some small degree of imprecision. These participants were therefore not excluded from analyses.

Table 1. Means and standard deviations for metacognitive measurement, trait inference error in the metacognition task, and covariates.

Variable	Mean	SD	Range
AQ score	62.96	8.96	41-83
TAS score	46.33	11.49	21-71
WASI-MR score	20.65	4.01	6-27
AUROC2	0.56	0.05	0.47-0.67
Mean absolute error in trait inference (Metacognition task)	0.87	0.14	0.63-1.19

Note. AQ = Autism Quotient, TAS = Toronto Alexithymia Scale, WASI-MR = WASI Matrix Reasoning, AUROC2 = area under the type 2 ROC (metacognitive sensitivity). *The WASI-MR should be interpreted as a proxy measure, and not a full measure of IQ. However, for our median age group (20-29 years), assuming approximately equal norm-referenced performance in the vocabulary and matrix reasoning components of the WASI FSIQ-2, a WASI-MR score of 21 would give an IQ estimate of 100.*

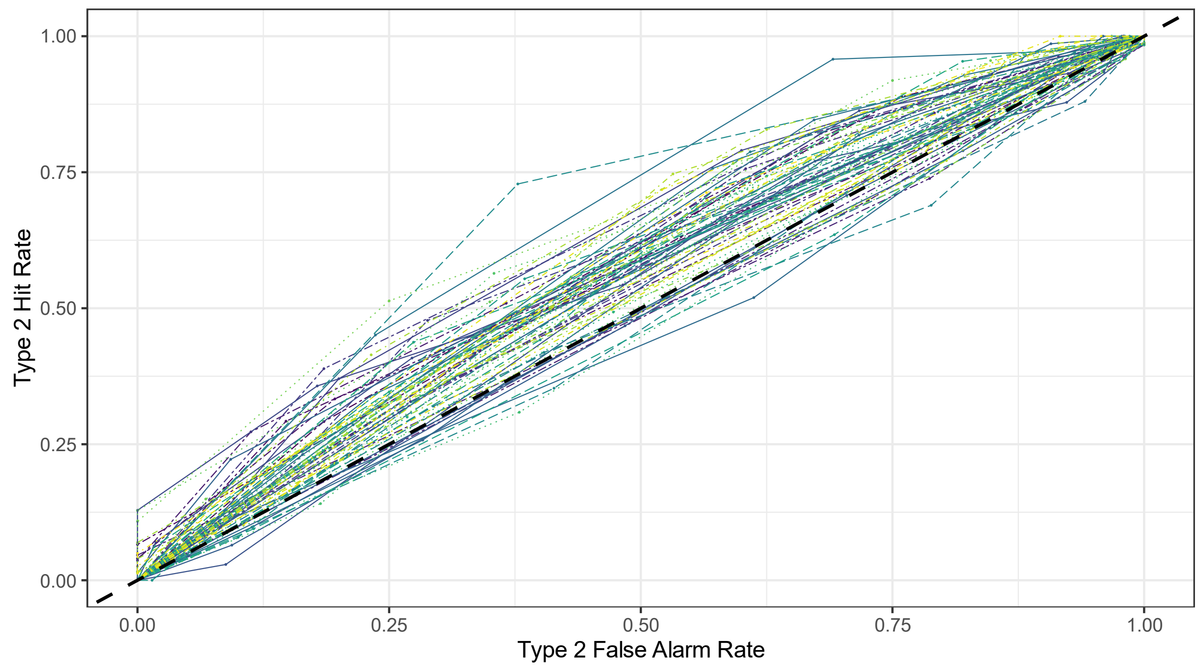

Figure 3. Plot of type 2 ROC curves. Each line is the curve of a single participant, and the thick dashed line represents chance metacognitive performance. Curves bowing to the top left of the line indicate better than chance performance, whilst curves bowing to the bottom right indicate worse than chance performance.

3.1. Theory of mind and metacognition

We conducted a linear multiple regression to determine whether metacognitive sensitivity was related to autistic traits, alexithymic traits, or intelligence. This regression found no association between participant AUROC2 scores and any of our covariate measures (all *p*s > .092). This regression model also contained participants' mean trait inference error in the metacognition task, in order to account for evidence that, theoretically, better first-order performance (in this case, reduced error) leads to increased AUROC2 values in the absence of higher metacognitive efficiency (Clarke et al., 1959; Galvin et al., 2003). In this case, however, we did not observe an association between metacognitive sensitivity (AUROC2) and first-order (i.e., trait inference) error on the metacognition task, *p* = .079. This model did not explain a significant amount of variance, *p* = .293. The effects of the covariates remained non-significant when mean trait inference error was excluded from the model (all *p*s > .093), and the reduced model did not explain a significant amount of variance (p = .404).

Descriptive statistics for trait inference error, mental state inference error, and reported confidence in the Interview Task are given in Table 2. We conducted a linear multiple regression testing our hypotheses that trait inference error propagates to produce error in mental state inference, and that better metacognitive sensitivity is associated with reduced error in mental state inference. This regression found a significant positive association between participant mean trait inference error in the Interview Task and participant mean mental state inference error, β = 0.51, *SE* = 0.12, t (75) = 4.38, p < .001, 95% CI [0.28, 0.74], but did not find an association between metacognitive sensitivity (AUROC2) score and participant mean mental state inference error, p = .208. The effect of trait inference error on mental state inference error is illustrated in Figure 4. There was no significant effect of mean trait inference error in the metacognition task on mean mental state inference error in the Interview Task, p = .124. The model explained a significant amount of variance F (3, 75) = 15.09, p < .001, R^2 = .38, R^2_{Adj} = .35. The relationship between AUROC2 score and participant mean mental state inference error remained non-significant when trait inference errors in both the Interview Task and metacognition task were removed from the analysis (p = .169).

Table 2. Means and standard deviations for trait inference error, mental state inference error, and reported confidence in the Interview Task.

Variable Level	Variable	Mean	SD	Range
Trial-by-trial	Trait inference error	0.83	0.62	0.00-2.70
	Mental state inference error	20.24	15.82	0-67
	Confidence report	3.57	0.95	1-5
Participant mean	Trait inference error	0.83	0.12	0.59-1.17
	Mental state inference error	20.28	2.33	14.70-28.20

Note. Trial-by-trial: given statistics were obtained from the raw values given on each trial of the Interview Task.

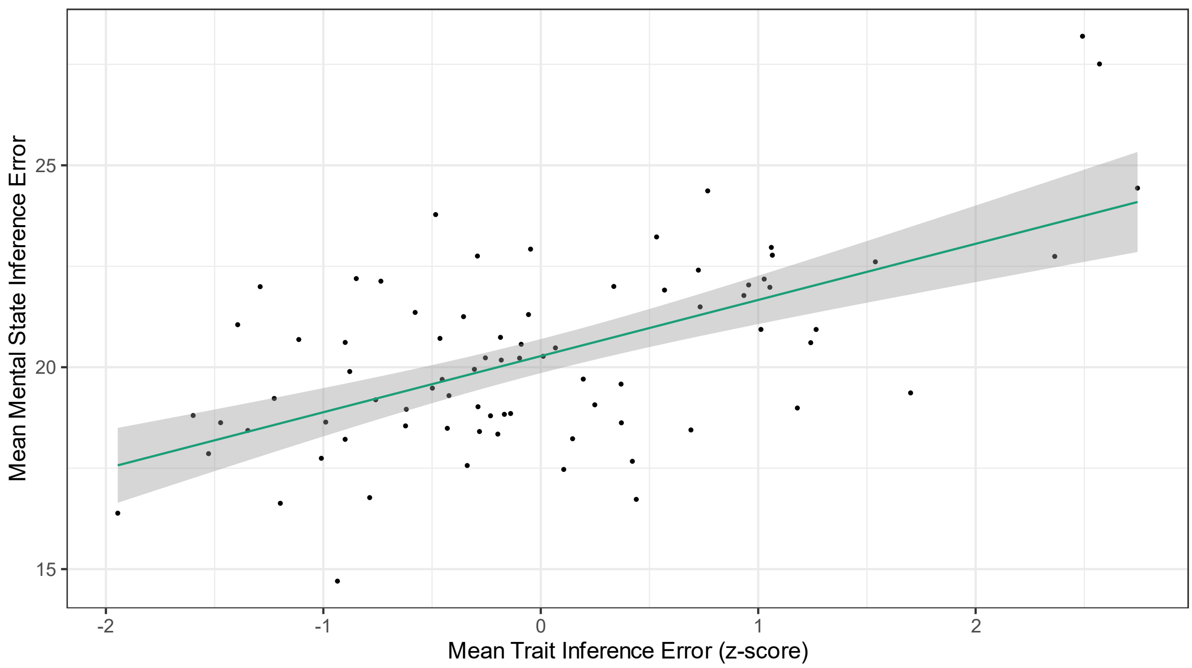

Figure 4. Relationship between mean trait inference error and mean mental state inference error in the Interview Task.

Our model testing the potential mechanism through which metacognition was hypothesised to influence mental state inference accuracy outperformed a null model including only the random intercepts of participant, video, and trial, as well as models including only the main effects of trait inference error in the Interview Task, AUROC2 and mean trait inference error in the metacognition task. The full model included, as predictors, participant error in a given trait inference for a given video stimulus, the reported confidence associated with this inference, the participant's metacognitive sensitivity and the participant's mean trait inference error in the metacognition task. Model comparison statistics are given in the Supplementary Materials (Section S.2.).

Table 3. Model summary for the best fitting model examining the potential mechanism for the metacognitive effect.

Random effects

Groups	Term	Variance	SD
Participant	Intercept	4.26	2.06
Trial	Intercept	15.62	3.95
Video	Intercept	5.06	2.25
Residual		225.31	15.01

Fixed effects

Term	Estimate	SE	df	t-value	p
Intercept	12.47	1.69	91.35	7.38	<.001
Trait Inference Error (Interview)	0.27	0.04	177455.14	7.10	<.001
Reported confidence	-0.23	0.05	99831.34	-4.53	<.001
AUROC2	-0.17	0.25	78.20	-0.70	.486
Trait Inference Error (Metacognition)	8.25	1.61	78.27	5.14	<.001
Trait Inference Error (Interview): Reported Confidence	0.08	0.04	177162.21	2.27	.023
Trait Inference Error (Interview): AUROC2	-0.00	0.04	176853.68	-0.12	.908
Reported Confidence: AUROC2	0.11	0.05	95557.46	1.98	.048
Trait Inference Error (Interview): Reported Confidence: AUROC2	-0.09	0.04	176960.22	-2.27	.023

All effects are reported in the model output given in Table 3. This model tested our prediction that the extent to which trait inference error is propagated to mental state inference error is determined by confidence, and that the greater coupling of error and confidence in individuals with higher metacognitive sensitivity would result, statistically, in a reduction of the two-way interaction effect, producing a three-way interaction effect. This expected three-way interaction effect between trait inference error in the Interview Task, the reported confidence in those inferences, and metacognitive sensitivity was significant (*B* = -0.09, *SE* = 0.04, *t* (176960.22) = -2.27, *p* = .023, 95% CI [-0.16, -0.01]). As shown in Figure 5, this interaction was such that confidence modulates the relationship between trait inference error and mental state inference error more strongly for those with low metacognitive sensitivity than those with high metacognitive sensitivity.

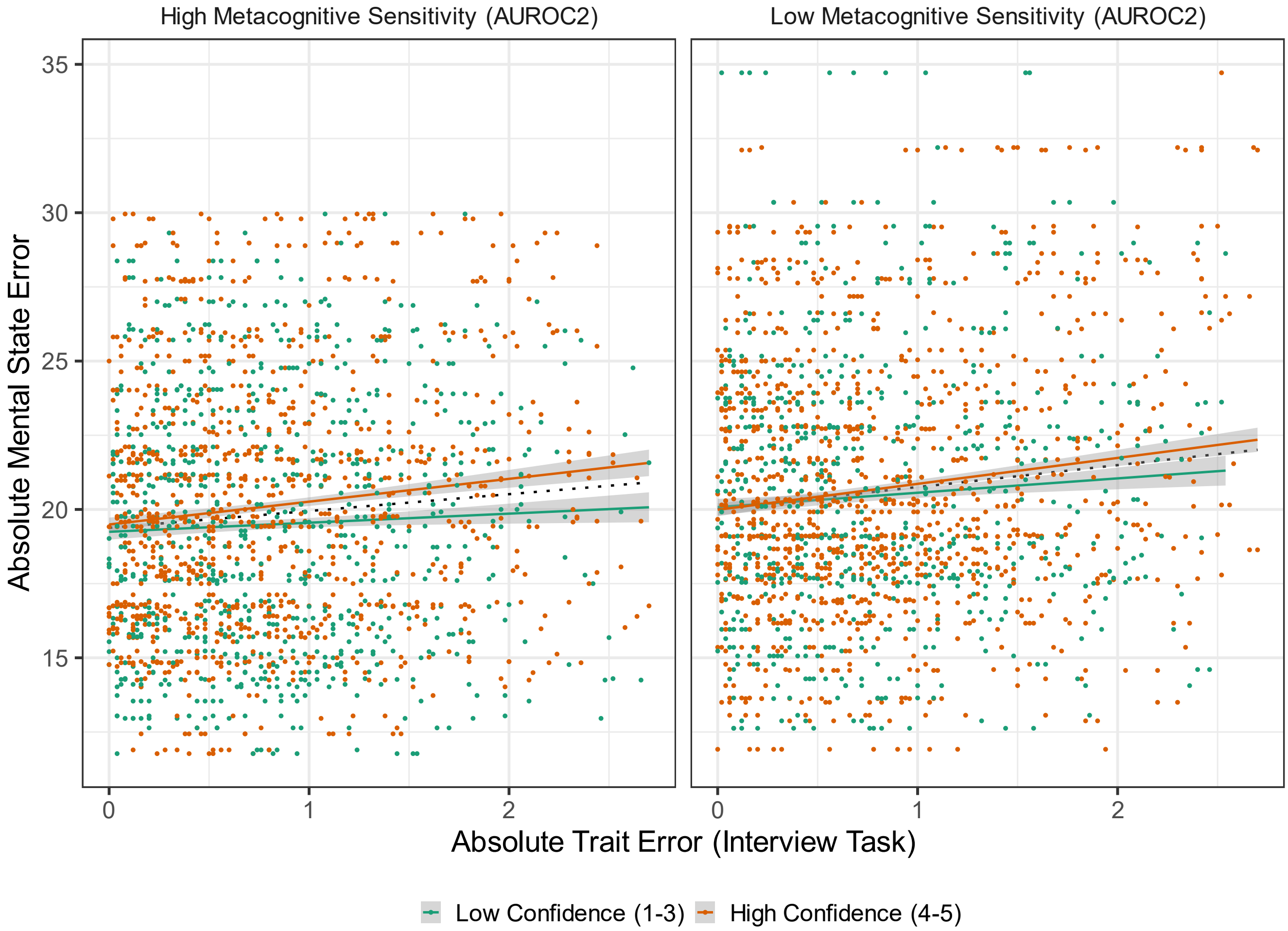

Figure 5. Three-way interaction between metacognitive sensitivity, confidence, and trait inference error when predicting mental state inference error. For individuals with low metacognitive sensitivity, confidence modulates the relationship between trait inference error and mental state inference error such that trait inference error is more positively related to mental state inference error when confidence is high, compared to when confidence is low. This effect is of smaller magnitude for individuals with high metacognitive sensitivity. For the purposes of these plots, 'High Metacognitive Sensitivity' is above sample median AUROC2, and 'Low Metacognitive Sensitivity' is below sample median AUROC2. *Note.* For the sake of visual interpretability, the y-coordinates of individual points in this figure represent the mean absolute mental state inference error across all mental states for a given video, with one point plotted for each trait judgement made regarding that video. The lines of best fit, however, are calculated from the full dataset used for modelling. This

dataset takes individual mental state judgements as separate datapoints. The shaded area represents standard error.

3.2. Cross-task comparisons

Cross-task analyses were conducted to test the assumption that (at least some of) the cognitive processes involved in trait inference and confidence formation in the metacognition task and the Interview Task are shared. We therefore predicted positive associations across the two tasks for each of our measures (i.e., mean trait inference error, participant-level correlation between confidence and error, and judgements of the traits of given target individuals).

Our first cross-task analysis showed a significant positive correlation between mean trait inference error in the metacognition task and mean trait inference error in the Interview Task, $r = .62$, $t(77) = 6.93$, $p < .001$, 95% *CI* [.46, .74]. Our second cross-task analysis showed a significant positive correlation between our proxy measures of metacognitive ability (the Pearson's correlation between trait inference error and confidence) across the two tasks, $r = .29$, $t(77) = 2.62$, $p = .010$, 95% *CI* [.07, .48].

When examining trials of the metacognition task which featured targets participants observed in the Interview Task, the model predicting Interview Task trait judgements on the basis of metacognition task trait judgements outperformed the null model. For this analysis, none of the possible random slopes explained a notable amount of variance, and so no random slopes model was included in this comparison. Model comparisons are given in the Supplementary Materials (Section S.3.).

Trait judgements made in the metacognition task significantly predicted trait judgements made in the Interview Task, $B = 3.18$, $SE = 0.54$, $t(1697.02) = 5.93$, $p < .001$, 95% CI [2.12, 4.27]. Full model statistics are given in Table 4.

Table 4*.* Model statistics for the association between trait judgements in the metacognition task and trait judgements in the Interview Task.

Random effects

Groups	Term	Variance	SD
Participant	Intercept	56.08	7.49
Trait	Intercept	24.75	4.98
Video	Intercept	8.04	2.84
Residual		298.75	17.28

Fixed effects

Term	Estimate	SE	df	t-value	p
Intercept	54.20	3.03	25.66	17.90	<.001
Trait Judgement (Meta)	3.18	0.54	1697.02	5.93	<.001

3.3. Predictors of trait inference accuracy

Descriptive statistics for error in perception of population median, error in self-perception, and participant-target difference are given in Table 5.

Table 5. Means and standard deviations for error in perception of population median, error in self-perception, and similarity.

Variable Level	Variable	Mean	SD	Range
Participant mean	Mean error in median perception	0.51	0.23	0.17-1.16
	Mean error in self-perception	0.74	0.28	0.21-1.55
Trial-by-trial	Participant-target difference (Meta)	0.83	0.61	0.00-3.70
	Participant-target difference (Interview)	0.80	0.60	0.00-3.70

Note. Trial-by-trial: given statistics were obtained from the raw values given on each trial of the metacognitive or Interview Task – this includes self-perception error on individual traits and absolute differences between participants and individual targets for individual traits. Participant mean: given statistics are reflective of the participants' mean error across all traits.

We predicted that participants who gave more erroneous estimates of population median traits would show greater error in trait inference. As predicted, trait inference error on the metacognition task was significantly positively associated with mean error in perception of median traits (β = 0.47, SE = 0.11, t (75) = 4.38, p < .001, 95% CI [0.26, 0.69]) but not with mean error in self-perception (p = .942). Together, these predictors explained a significant amount of variance in trait inference error on the metacognition task, F (2,75) = 10.75, p < .001, R^2 = .22, R^2_{Adj} = .20. The same effects were observed for trait inference error on the Interview Task, where we observed a significant positive association with mean error in perception of median traits (β = 0.42, SE = 0.11, t (75) = 3.78, p < .001, 95% CI [0.20, 0.64]), but not with mean error in self-perception (p = .677). Again, this model explained a significant amount of variance, F (2, 75) = 8.83, p < .001, R^2 = .19, R^2_{Adj} = .17. These effects are illustrated in Figure 6.

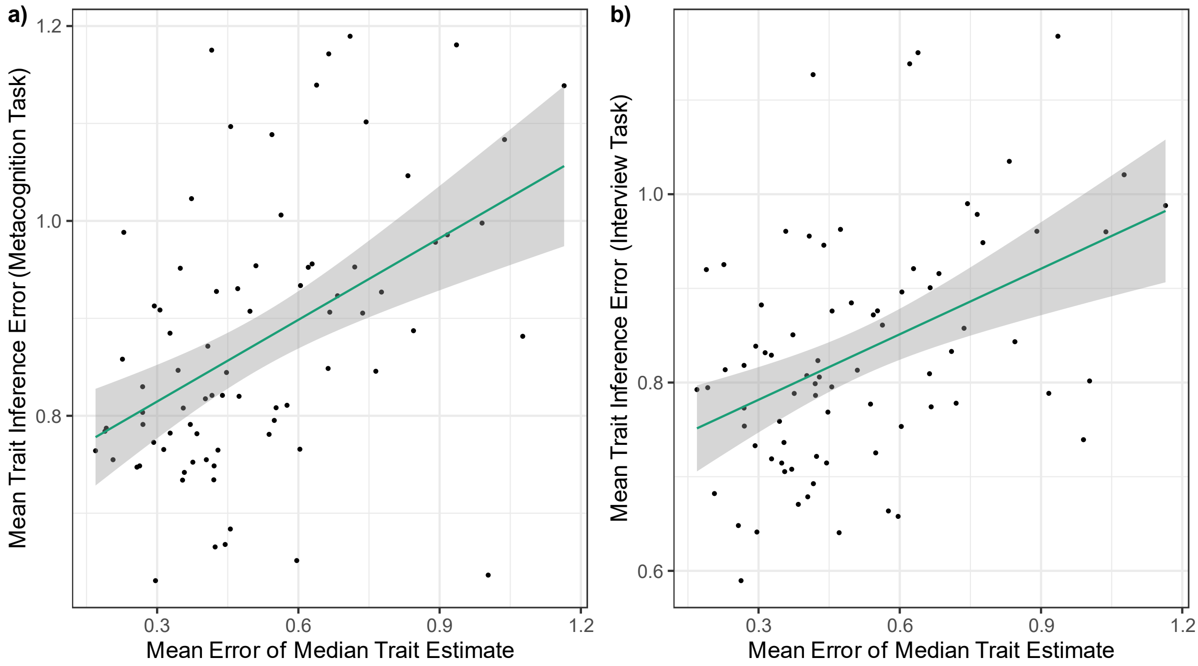

Figure 6. a) Relationship between participant mean error in estimates of median population trait values and participant mean trait inference error on the metacognition task. b) Relationship between participant mean error in estimates of median population trait values and participant mean trait inference error on the Interview Task.

Mental state inference error in the Interview Task was also positively associated with mean error in perception of median traits (β = 0.31, *SE* = 0.12, *t* (75) = 2.64, *p* = .010, 95% *CI* [0.08, 0.54]) but not with mean error in self-perception (*p* = .860). Again, this model explained a significant portion of the variance, *F* (2, 75) = 4.16, *p* = .019, R^2 = .10, R^2_{Adj} = .08. However, the association between error in perception of median traits and mental state inference error was not observed when trait inference error on the Interview Task was included in the analysis (*p* = .511). Echoing our earlier finding, mental state inference error in the Interview Task was positively associated with trait inference error in the Interview Task, β = 0.57, *SE* = 0.10, *t* (75) = 5.58, *p* < .001, 95% *CI* [0.37, 0.77]. This model explained a large portion of the variance in mental state inference error, *F* (2, 75) = 21.40, *p* < .001, R^2 = .36, R^2_{Adj} = .35.

We conducted linear mixed effects modelling to test our hypothesis that participants would make more accurate trait inferences for participants who were more similar to them, but that this effect would be modulated by the accuracy with which participants perceived their own traits. Specifically, participants who showed greater self-perception error were expected to gain less benefit from similarity, such that the increase in the error of trait inference with increasing participant-target difference would be smaller in magnitude than for participants with lower self-perception error.

For models examining the associations of participant-target similarity and participant self-perception error with trait inference error in the metacognition task, the best fitting random slopes model allowed the slope of participant-target difference to vary as a function of participant, but not trait or video stimulus. For models predicting trait inference error in the Interview Task, the best fitting model allowed the slope of participant-target difference to vary as a function of participant and trait, but not video stimulus. In both cases, the random slopes model outperformed the null model, a model including participant-target difference only, and the intercepts-only model. Full model comparisons are given in the Supplementary Materials (Section S.4.).

When predicting trait inference error on the metacognition task, we observed a significant positive association with participant-target trait difference ($B = 0.11$, $SE = 0.01$, $t(78.34) = 8.33$, $p < .001$, 95% CI [0.08, 0.13]) and a significant interaction effect ($B= -0.03$, $SE = 0.01$, $t(2801.03) = --4.30$, $p < .001$, 95% CI [-0.04, -0.02]). As illustrated in Figure 7, the interaction effect was such that the similarity effect (i.e., reduced trait inference error for targets who are more similar to the participant) was reduced for those who showed greater error in self-perception. Full model statistics are provided in Table 6.

Table 6. Model statistics for the association between trait inference error in the metacognition task and participant-target trait difference and participant self-perception error.

Random effects				
Groups	Term	Variance	SD	Correlation
Participant	Intercept	0.02	0.13	
	Trait difference	0.01	0.10	-.19
Video	Intercept	0.01	0.12	
Trait	Intercept	0.01	0.12	
Residual		0.38	0.62	

Fixed effects					
Term	Estimate	SE	df	t-value	p
Intercept	0.89	0.06	10.73	15.31	<.001
Trait difference	0.11	0.01	78.34	8.33	<.001
Self-perception error	0.01	0.01	7773.30	1.45	.148
Trait difference: self-perception error	-0.03	0.01	2801.03	-4.30	< .001

The same pattern of results was observed when predicting trait inference error on the Interview Task. A significant positive association between participant-target trait difference and trait inference error was observed (B = 0.11, SE = 0.02, t (11.47) = 4.41, p < .001, 95% CI [0.06, 0.15]), as well as a significant interaction effect (B = -0.08, SE = 0.01, t (1850.14) = -9.58, p < .001, 95% CI [-0.10, -0.07]). As shown as Figure 7, the interaction effect was once again such that the similarity effect was reduced for those who showed greater error in self-perception. Full model statistics are provided in Table 7.

Table 7. Model statistics for the association between trait inference error in the Interview Task and participant-target trait difference and participant self-perception error.

Random effects

Groups	Term	Variance	SD	Correlation
Participant	Intercept	0.01	0.11	
	Trait difference	0.01	0.12	.03
Video	Intercept	0.00	0.05	
Trait	Intercept	0.00	0.06	
	Trait difference	0.00	0.04	.41
Residual		0.34	0.59	

Fixed effects

Term	Estimate	SE	df	t-value	p
Intercept	0.85	0.03	13.37	26.86	<.001
Trait difference	0.11	0.02	11.47	4.41	<.001
Self-perception error	0.01	0.01	3958.84	1.50	.134
Trait difference: self-perception error	-0.08	0.01	1850.14	-9.58	< .001

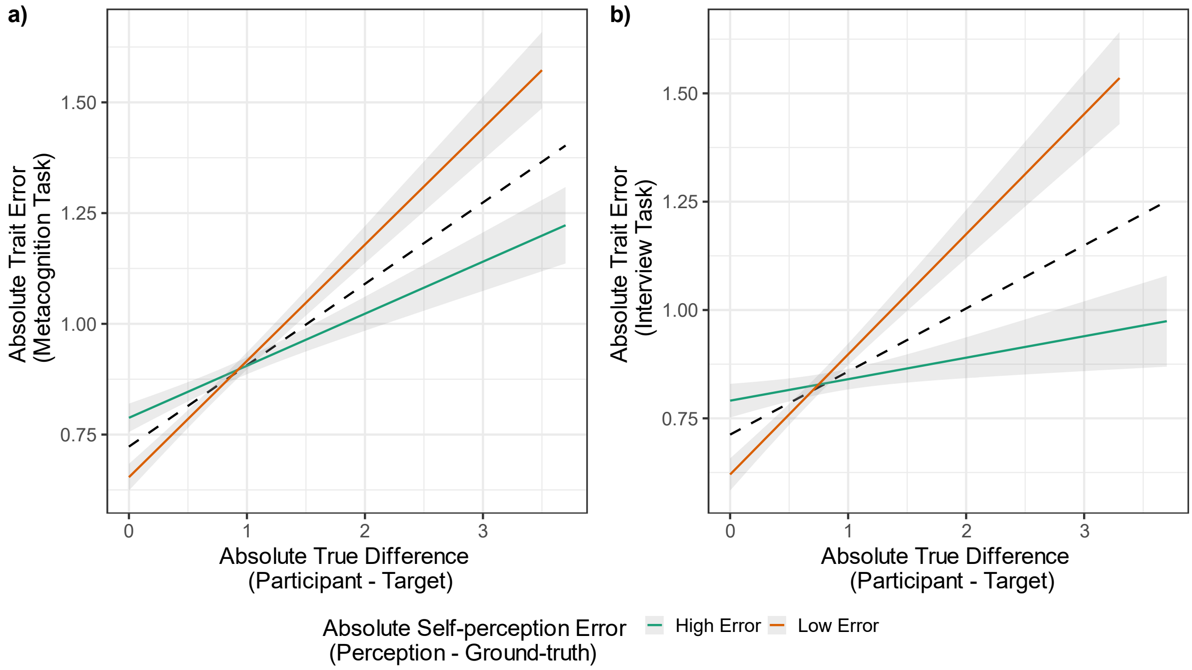

Figure 7. a) Two-way interaction between participant-target trait difference and participant self-perception error in predicting trait inference error in the metacognition task. b) Two-way interaction between participant-target trait difference and participant self-perception error in predicting trait inference error in the Interview Task. In both cases, the positive relationship between participant-target trait difference and trait inference error is reduced in participants who show greater error in self-perception. For the purposes of these plots, 'High Error' is above sample median error in self-perception, and 'Low Error' is below sample median error in self-perception. The dotted line shows the overall effect across both groups. Shaded areas represent standard error.

3.4. Predictors of confidence in trait inference

To test our hypothesis that participants would be more confident in trait judgements regarding individuals that they perceive to be more similar to them, and that this confidence would be misplaced in individuals with poor awareness of their own traits, we fitted linear mixed effects models. These models examined the extent to which participants' reported confidence in trait inferences made during the metacognition task could be predicted by the error of the inference in question, the difference between the participant and the target on the trait being inferred, the error in the participant's perception of themselves on that trait dimension, and the interactions between these predictors. As discussed in the Introduction, we predicted a three-way interaction, such that the relationship between trait inference error and confidence would be more positive (i.e., participants would be more confident in erroneous inferences) when participant-target difference was low and participant self-perception error was high.

Here we report results from linear models fitted with confidence treated as a continuous variable, but it should be noted that the same results are observed using equivalent models fitted on confidence as an ordinal variable – these models are reported in the Supplementary Materials (Section S.5.). The best fitting random slopes model allowed the slope of trait inference error, but not participant-target difference or participant self-perception error, to vary by trait, video stimulus, and participant. This model outperformed the null model, models containing only the main effects and interactions of trait inference error and participant-target difference, and the intercepts-only model. Full model comparisons are given in the Supplementary Materials (Section S.5.).

As predicted, we observed that confidence in trait inferences was significantly associated with a three-way interaction between trait inference error, participant-target difference in the relevant trait, and participant self-perception error in that trait (B = -0.04, SE = 0.01, t (9745.43) = -6.41, p < .001, 95% CI [-0.05, -0.03]. As illustrated in Figure 8, this interaction effect was such that participant confidence judgements were more positively related to trait inference error (i.e., participants were more confident in *less* accurate trait judgements) when targets were similar to them, and this effect was greater in participants with more erroneous perceptions of their own traits. Full model statistics are provided in Table 8.

Table 8. Model summary for the best fitting model examining predictors of confidence in trait inference in the metacognition task. for the association between confidence in trait inferences in the metacognition task and trait inference error, participant-target trait difference and participant self-perception error.

Random effects

Groups	Term	Variance	SD	Correlation
Participant	Intercept	0.38	0.62	
Video	Intercept	0.01	0.11	
Trait	Intercept	0.01	0.10	
Residual		0.61	0.78	

Fixed effects

Term	Estimate	SE	df	t-value	p
Intercept	3.55	0.08	57.85	42.04	<.001
Trait inference error	0.05	0.01	9740.59	6.23	<.001
Trait difference	-0.01	0.01	9749.82	-0.79	.428
Self-perception error	0.00	0.01	9779.51	0.07	.943
Trait inference error: trait difference	-0.03	0.01	9752.66	-3.86	<.001
Trait inference error: self-perception error	0.02	0.01	9737.56	2.81	.005
Trait difference: self-perception error	0.01	0.01	9761.46	1.31	.191
Trait inference error: trait difference: self-perception error	-0.04	0.01	9745.43	-6.41	<.001

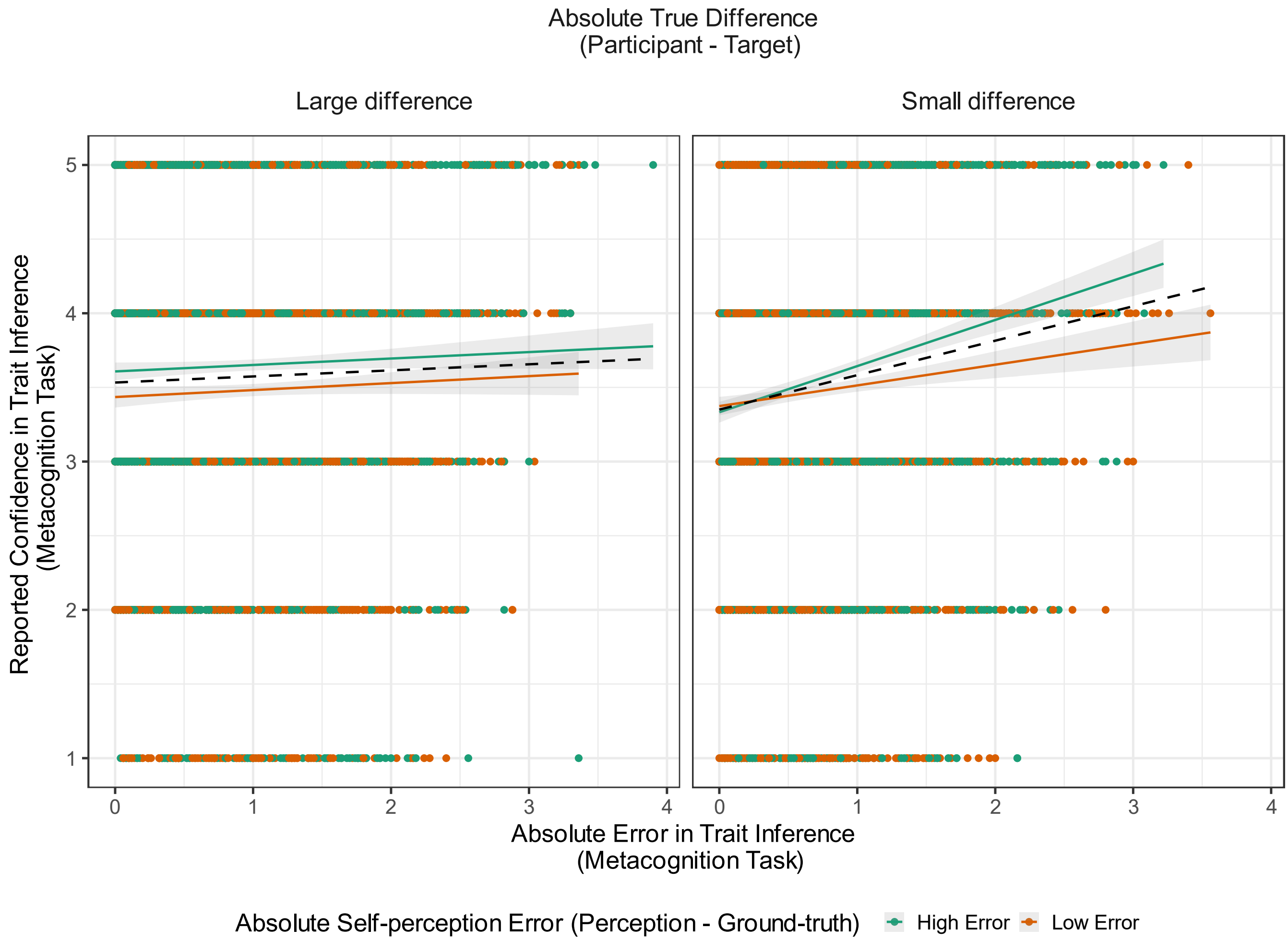

Figure 8. Three-way interaction between error in trait inference, participant-target trait difference, and participant self-perception error in predicting reported confidence in trait inferences made in the metacognition task. Participants are more likely to be more confident in erroneous trait inferences when the target is similar, rather than dissimilar, to them. This effect is greater in participants with inaccurate self-perception. For the purposes of this plot, 'Large difference' is above sample median absolute difference in HEXACO trait score between the participant and the target (i.e., the target is dissimilar to the participant), and 'Small difference' is below sample median absolute trait difference (i.e., the target is similar to the participant). Additionally, 'High Error' is above sample median error in self-perception, and 'Low Error' is below sample median error in self-perception. The dashed lines represent the overall relationship between confidence and trait error

for dissimilar and similar targets across degrees of self-perception error. Shaded areas represent standard error.

4. Discussion

This study sought to identify the mechanisms underlying ToM inference. To do so, we used a novel metacognition task in which metacognitive sensitivity in the domain of trait inference could be quantified. We also used the Interview Task, a ToM task in which ground-truth information is available, to assess the accuracy of participants' inferences regarding targets' traits and mental states.

Our first set of analyses tested predictors of metacognitive sensitivity and ToM ability. We observed no association between metacognitive sensitivity and autistic traits, alexithymic traits or intelligence. We also found no significant association between metacognitive sensitivity and participant mean error of ToM inferences in linear multiple regression. However, a significant three-way interaction between metacognitive sensitivity, Interview Task trait inference error and confidence in trait inference suggested that confidence modulates the relationship between errors in trait inference and errors in mental state inference, but that this interaction is smaller in magnitude in participants with higher metacognitive sensitivity.

Our second set of analyses, testing possible predictors of trait inference accuracy, demonstrated that in both short (30 second) and longer (four minute) videos, participants who showed a more accurate understanding of population median traits also showed reduced error in their inferences about targets' traits. Furthermore, participants showed reduced error in trait inferences for targets who were more similar to them, but this similarity effect was modulated by the accuracy with which participants perceived their own traits, such that participants with less accurate self-perception gained less benefit from target similarity. Again, these effects were observed in both the shorter videos of the metacognition task and the longer videos of the Interview Task.

Our final set of analyses, testing predictors of confidence in trait inference, revealed a significant three-way interaction between trait inference error, participant-target similarity, and participant self-perception error. This effect was such that the relationship between trait inference error and confidence was more positive when participants and targets were more similar to one another, and this two-way effect was heightened when the participant's estimate of their own traits was more erroneous.

4.1. Theory of Mind and metacognition

The study reported here brings novel insights into the process of mental state inference by providing evidence for an explanation of the relationship between metacognition and ToM ability that is not considered by the primary theories linking the two abilities (Carruthers, 2009, 2011; Carruthers & Smith, 1996; Goldman, 2006; Nichols & Stich, 2003). Namely, we suggested that metacognitive ability may be useful in weighting trait inferences to optimise the accuracy of mental state inferences. We predicted that when participants reported higher confidence in a trait inference, any error in that trait inference would be more likely to be propagated into associated mental state inferences. As such, the relationship between trait inference error and mental state inference error was expected to be stronger when confidence is high, as more of the error in trait inference is propagated to the mental state inferences than when confidence is low.

We hypothesised that with higher metacognitive sensitivity, indicating a better ability to discriminate between accurate and inaccurate trait inferences, high confidence trait inferences would (by definition) be more accurate, and thus there should be less error to be propagated to the mental state inferences. Furthermore, error from low confidence trait inferences, which would be less accurate, will be less likely to be propagated; instead, the mental state inferences will be determined by other available information, including other more accurate trait inferences. As such, an individual with high metacognitive sensitivity should use trait inferences more optimally, such that mental state inferences are as accurate as possible given the available information. Based on this, we predicted that the strong coupling of trait inference error and reported confidence would reduce the magnitude of the two-way interaction between trait inference error and confidence in mental state inferences. In contrast, the decoupling of confidence from trait inference error in participants with lower metacognitive sensitivity means that the two-way interaction should be larger, because the trait inference error that may or may not be propagated is more evenly distributed across levels of reported confidence. This statistical pattern is to be expected due to the level of coupling between error and confidence but does not imply that there is a functional difference in the use of trait information and confidence in individuals with differing levels of metacognitive sensitivity. Instead, this interaction demonstrates that the hypothesised weighting process results in differential outcomes dependent on an individual's awareness of the accuracy of their trait judgements.

To further illustrate, we can take each case in turn. When there is little error in a trait inference, an individual with high metacognitive sensitivity would be very likely to be confident in that inference and therefore should lend it substantial weight in determining the mental state inference. That small amount of error will therefore be passed on into the mental state inference. When there is a lot of error in a trait inference, an individual with high metacognitive sensitivity will usually recognise this and will therefore put little reliance on (or entirely discard) that trait inference, meaning that the error in this trait inference will not be passed to the mental state inference. In this case, the statistical relationship between trait inference error and confidence is high (because the mentaliser is sensitive to the accuracy of their inference and this is reflected in their confidence). Statistically, this close relationship between trait inference error and confidence decreases the modulatory effect that confidence would be expected to have on the relationship between trait inference error and mental state inference error. This is because much of the variance in confidence is shared with variance in trait inference error, such that there are relatively few trials in which a low confidence rating is given to an accurate judgement, or a high confidence rating is given to an inaccurate judgement. As such, this statistical effect is reflective of how metacognitive sensitivity facilitates optimal weighting of trait inferences.

In contrast, individuals with low metacognitive sensitivity are less able to discriminate between accurate and inaccurate trait inferences. Therefore, we would expect that an individual with low metacognitive sensitivity will be more likely to have low confidence in accurate trait inferences (and thus down-weight or discard useful inferences) and to have high confidence in inaccurate trait inferences. When there is a lot of error in an inference, they may, therefore, lend this trait inference substantial weight in determining the mental state inference, resulting in a large amount of error being passed to the mental state inference. In this case, because the individual is less able to discriminate between accurate and inaccurate trait inferences, the statistical relationship between trait inference error and confidence is smaller. As such, there is a larger proportion of variance in confidence that is not shared with variance in trait inference error. This means that the modulatory effect of confidence on the relationship between trait inference error and mental state inference error can be more readily observed statistically. Therefore, whilst the same process of weighting trait inferences according to confidence is thought to occur across all levels of metacognitive sensitivity, differences in the relationship between trait inference accuracy and confidence across different levels of metacognitive sensitivity means that this is statistically observed as a three-way interaction.

As this predicted three-way interaction was observed, our results indicate that metacognition plays a role in the use of trait information in mental state inference. However, we did not find a significant association between AUROC2 (our measure of metacognitive ability) and participant mean mental state inference accuracy in multiple linear regression. We suggest a possible explanation for this pattern of results in Section 4.3, considering all findings from the present study.

Given that the AUROC2 measure obtained through our metacognition task was used to examine the use of trait information in the Interview Task, it was important to examine whether individual differences in trait inference and related confidence judgements diverged across our two tasks. Specifically, we wanted to test the assumption that the process of making these judgements based on relatively little information (in our shorter metacognitive videos) was related to the process of making the same judgements on the basis of more information (in our longer Interview Task videos). The observed associations between judgements and performance across tasks are therefore supportive of the idea that our metacognitive measure validly captures ability in the metacognitive process of interest, especially given that participants made substantially fewer trait inferences and confidence judgements in the Interview Task, and thus accuracy-confidence correlations are less likely to be stable in the Interview Task. However, as we will discuss in Section 4.3., it should be noted that our analysis of confidence reports in the metacognition task indicates that there may be target-specific, within-participant differences in metacognitive sensitivity. As such, our metacognition measure should not be considered a pure measure of an 'overall' metacognitive sensitivity in the trait inference domain.

Ultimately, then, the present study provides evidence for a mechanism through which metacognition can influence ToM. However, further work is required to assess the extent to which our proposed mechanism may explain previously observed associations involving performance in other ToM tasks (K. L. Carpenter et al., 2019; Nicholson et al., 2020; van der Plas et al., 2021; D. M. Williams et al., 2018). In particular, the most common ToM tasks used to test these associations may, in logical terms, have a less clear mechanistic role for metacognitive ability. Neither the Reading the Mind in the Eyes Test nor the Frith-Happé Animations Test have a direct trait inference component – participants are not explicitly required to make or use trait inferences about the targets of their mental state inferences. Therefore, it is possible that previously observed associations between metacognition and performance in these tasks may occur through some other mechanism to that discussed in the present paper.

However, it is also possible, given the naturalistic character of the Interview Task, that the mechanism described here underlies an association between metacognitive ability and ToM ability in day-to-day life, and that this relationship has downstream effects on more constrained experimental tasks. For example, the Frith-Happé Animations Test assesses the extent to which participants tend to make accurate mentalistic inferences about shapes. This task therefore tests both the accuracy of participants' inferences (albeit relative to an experimenter-defined standard, rather than ground-truth) and participants' propensity to make such inferences. It may be that individuals with poorer metacognitive ability tend to make less accurate mental state inferences based on trait information in everyday life and, because of this, show a reduced propensity to make mental state inferences at all, as the inferences they make are often of limited value in predicting or explaining behaviour. Similarly, if poor metacognitive ability leads to diminished mental state inference accuracy in day-to-day life, participants may have a worse understanding of mental states even without the context of traits. That is, if their mental state inferences are often less accurate, then they will be less able to draw conclusions about the 'average' mental states (across different locations in Mind-space) that may be represented in the Frith-Happé Animations Test. Further work is required to test these ideas and examine exactly how, if at all, different ToM tasks functionally relate to one another.

Whilst this study did provide novel insights into the relationship between ToM and metacognition, it may not conclusively contribute to the debate as to whether autism is characterised by a metarepresentational deficit that causes difficulties with metacognition and ToM. As noted in the Introduction, we had no prior predictions regarding the relationship between our covariates (most notably autistic traits, as measured by the AQ) and metacognition. Much of the body of evidence that might lead one to expect a negative association between autistic traits and metacognitive sensitivity examined group differences between diagnosed autistic participants and neurotypical participants (Grainger et al., 2016; Nicholson et al., 2020; van der Plas et al., 2021; D. M. Williams et al., 2018; Wojcik et al., 2013) and these group differences have not always been observed (K. L.

Carpenter et al., 2019; Wilkinson et al., 2010; Wojcik et al., 2011; Wojcik et al., 2013). An association between metacognitive ability and AQ score has previously been observed (K. L. Carpenter et al., 2019), but at least two studies have failed to find this association (van der Plas et al., 2021; D. M. Williams et al., 2018).

One possible explanation for this mixed literature, and for our own null finding regarding the association between metacognition and autistic traits, lies in the question of whether the AQ validly measures differences that may affect metacognitive ability. Specifically, measuring autistic traits as a continuous property in a neurotypical population may give different results to comparing neurotypical participants to those with a diagnosis of autism. Whilst there is a body of evidence suggesting that autistic traits are normally distributed across the population and that those who meet diagnostic thresholds for autism are at the extreme end of that distribution (Constantino & Todd, 2003; Ruzich et al., 2015) there are also questions surrounding whether continuous measures such as the AQ are valid predictors of autism diagnosis (Ashwood et al., 2016; Sizoo et al., 2015) and therefore whether continuously measured autistic traits are qualitatively, not just quantitatively, different from the pattern of symptoms observed in autism.

A second possible explanation lies in the methodology of this study relative to other studies. Our measure of metacognitive sensitivity was independent of metacognitive confidence, a feature that has, to our knowledge, been present in only two other studies examining metacognition and ToM (Nicholson et al., 2020; van der Plas et al., 2021). Furthermore, our study is the first to examine metacognitive ability specifically in the domain of trait inference, rather than perception (K. L. Carpenter et al., 2019; Nicholson et al., 2020; van der Plas et al., 2021), knowledge (D. M. Williams et al., 2018), or memory (Grainger et al., 2014, 2016; Wilkinson et al., 2010; Wojcik et al., 2011; Wojcik et al., 2013). There is evidence to suggest that average confidence in task performance differs between autistic and neurotypical individuals (McMahon et al., 2016; Z. J. Williams et al., 2022; Zalla et al., 2015), as well as evidence of group differences in sensory sensitivity (which may affect first-order perceptual performance) (Ashwin et al., 2009; Jussila et al., 2020; Takarae et al., 2016) and memory (Griffin, Bauer, & Gavett, 2022; Southwick et al., 2011; D. L. Williams, Goldstein, & Minshew, 2006). Therefore, it is possible that these more general cognitive differences, rather than deficits in metacognitive ability itself, underlie previously observed differences in measured metacognitive performance between autistic and non-autistic participants. If it is the case that ToM and metacognition are subserved by a single system that is damaged in autism, measuring metacognitive ability in a domain known to be directly relevant to mental state inference (Conway et al., 2020; Long et al., 2022) should theoretically maximise the chance of finding an association between autistic traits and metacognitive ability (and also between ToM and metacognitive ability). Similarly, our sample included a broad range of scores on both the AQ and TAS (measures of autistic and alexithymic traits, respectively), and several participants scored above threshold on either or both measures, suggesting that our null result is not a product of a limited range of either set of traits. Therefore, due to the absence of this finding, we find no evidence that autistic traits (albeit possibly distinct from a diagnosis of autism) are the result of dysfunction in a single metarepresentational system.

In addition to testing hypotheses regarding the relationship between metacognition and ToM, the present study provided a replication of the finding that trait inference error is associated with mental state inference error in the Interview Task (Long et al., 2022). The first study using the Interview Task utilised analyses in which each trait inference was considered separately and shown to have differential directional relationships with specific mental state inferences. In contrast, this study made use of the mean of the absolute error of participants' trait and mental state inferences. The result of this higher-level analysis demonstrates that the Interview Task provides a sensitive measure of both trait inference and mental state inference accuracy, and further supports the central tenet of the Mind-space theory: that trait inference underpins, to some extent, mental state inference. Future work should seek to examine the reliability of the Interview Task in detecting stable individual differences in ability should they exist.

These studies cannot, however, give a full picture of the dynamics of the relationship between trait representation and mental state inference. It is logical that one would make use of information regarding stable characteristics of individuals (i.e., traits) to infer momentary mental states. Indeed, evidence that the relationship between trait and mental state inference is modulated by confidence in specific trait inferences, presented in this paper, supports this notion. However, it is also plausible that if a mentaliser receives feedback about an inaccurate mental state inference, resultant prediction error might lead to an update in their representation of the target, either in terms of the target's location on particular trait dimensions, or in terms of the dimensions on which that target is represented.

As discussed in the Introduction, the Interview Task measures the accuracy of mental state inferences against ground-truth information obtained from the target of inference, rather than an experimenter- or consensus-defined standard. It is important, therefore, to consider whether self-reported mental states can truly be considered 'ground-truth'. Whilst the use of self-report leaves open the possibility of target participants misreporting their mental states, there is no clear reason to expect them to do so. It was made clear that responses would not be shown to the participant's interview counterpart, and questions tended not to have one response that would be more socially desirable than another. As such, there was no incentive to respond in a particular manner in this task and, furthermore, giving honest answers could help the participant to improve their interview ability based on the practice interview.

Even in the absence of intentional misreporting, one might suspect that individuals could lack awareness of the mental states underlying their actions. It is certainly possible that some individuals may be poor at predicting their future mental states, recalling past mental states, or predicting their behaviour based on their mental states. In contrast, the attitude one holds towards a particular proposition at a given moment (e.g., whether one *currently* believes that the candidate is performing well in the interview) can necessarily (only) be accessed by oneself at that time (Gertler, 2010).

Similarly, even if certain propositional content was not evaluated prior to the participant being asked to consider that proposition, upon prompting the resultant propositional attitude is necessarily that individual's mental state.

Reports of current propositional attitudes, then, in the absence of intentional misreporting, should be considered as true ground-truth mental states. It should be noted that this would not be the case for a retrospective paradigm in which target individuals recall past events and their mental states during these events, as memory is highly malleable (Bartlett, 1932; Maehara & Umeda, 2013) and the target would thus need to reconstruct or infer their previous mental states based on stored information, rather than accessing them directly. A predictive paradigm in which target individuals report what their mental state *would* be in a given situation would be similarly limited, as future mental states are also not directly accessible and would need to be inferred based on self-knowledge. As such, the use of ground-truth reports of targets' *current* mental states (at the time of reporting) is an important, and substantially beneficial, feature of the Interview Task.

4.2. Predictors of trait inference accuracy

In seeking to explore mechanisms underlying mental state inference, the present study also examined possible predictors of trait inference accuracy, which is itself known to be associated with mental state inference accuracy (Long et al., 2022). We found the same pattern of results across both our shorter and longer video stimuli, again suggesting that trait inference based on relatively little information relies on the same processes as trait inference based on more substantial information.

As predicted, trait inference accuracy was associated with the accuracy with which our participants perceived the 'average' mind. It is plausible that the process of trait inference involves evaluating targets against the population average, akin to the norm-based model of Face-space (Mueller, Utz, Carbon, & Strobach, 2020; Valentine, Lewis, & Hills, 2016; Wuttke & Schweinberger, 2019). However, there are alternative explanations that also may account for this effect. Specifically, participants who are better able to report the population median of a trait dimension may be able to do so because they have experienced a more representative sample of individuals across their lifetime. Mind-space theory would predict that these participants have a more accurate Mind-space (i.e., they will be better able to represent population covariance between dimensions) and that they would be more familiar with how different behavioural presentations correspond to Mind-space location. According to both predictions, these participants would therefore be expected to be better at locating specific targets in Mind-space, as we observed in this study. Another potential explanation may be that participants who are better at locating individuals in Mind-space are better able to intuit the population median because they have accurate data on which to base their judgement. A participant who has experienced a representative sample of the population, but routinely mis-locates individuals in Mind-space, would be unable to accurately infer the population median value, as they would be taking the median of erroneous trait inferences. In practice, it is likely that both factors may be at play here.

It is worth noting that the sample used to obtain the population median was a Canadian student sample (Ashton & Lee, 2009). There are sizeable differences in average scores on the HEXACO-60 dimensions between student and community samples (Ashton & Lee, 2009; Lee & Ashton, 2018) and so it is arguably more correct to say that those who were more accurate in their perception of the student population median were more accurate in trait inference in the Interview Task. However, the majority of targets in our Interview Task stimuli were themselves students, and so one would expect an accurate understanding of the student population median to be more useful in this case than an accurate understanding of the broader population median. From the present data, then, we cannot be certain that those who gave accurate reports of the median are likely to be better at trait inference when the targets are representative of the general population. Nevertheless, given the consistency between the sample used to obtain the median and the sample of targets, this evidence suggests that there is an association between the accuracy of one's understanding of the median traits of the target population and the accuracy of trait inferences regarding members of that population.

Whilst cross-cultural differences between populations in Canada and in the UK might have influenced the measured accuracy of participants' perceptions of median traits, previous studies have shown that mean values of HEXACO traits across these two countries differ less than between student and community populations (Ashton & Lee, 2009; Lee & Ashton, 2018; Lee, Ashton, Griep, & Edmonds, 2018). Furthermore, there is no reason to suspect that individuals who are relatively more attuned to Canadian than British minds would perform better when estimating the traits of our targets. As such, any influence of cross-cultural differences on the measurement of participants' perceptions of median traits is unlikely to affect the conclusions of this study. However, further research is needed to ensure that this is the case – such research should assess participants' understanding of population median traits using ground-truth data obtained from a sample which is culturally congruent with the population from which the targets of trait inference are sampled.

We additionally replicated the previously observed similarity effect (Conway et al., 2020), in which participants are more accurate at locating individuals in Mind-space when that individual is more similar to them. The present study also demonstrated that, as predicted by the Mind-space theory (Conway et al., 2019), the similarity effect is modulated by the accuracy of the participant's self-perception. This interaction is expected because, given a participant is more likely to recognise behaviour that is similar to their own and thus successfully locate the target as occupying a similar space to them in Mind-space, if they represent their own location in Mind-space inaccurately, this inaccurate location is also attributed to the target. A possible limitation of this study in examining this effect is the use of self-report personality questionnaires to measure participants' and targets' 'true' traits.

The HEXACO-PI-R has been shown to have high reliability and high agreement between self- and other-reports (Moshagen, Thielmann, Hilbig, & Zettler, 2019). It is also known to be less susceptible to social desirability bias than other personality questionnaires (Lee & Ashton, 2013). HEXACO self-report measures have also been shown to have strong predictive validity for both reported and observed behaviour in a variety of domains, including prosocial behaviour (Thielmann, Spadaro, & Balliet, 2020), unethical behaviour (Heck, Thielmann, Moshagen, & Hilbig, 2018), popularity and likeability (de Vries, Pronk, Olthof, & Goossens, 2020), and pro-environmental attitudes and behaviours (Soutter, Bates, & Mõttus, 2020). As such, it is highly likely that self-report responses to the HEXACO-60 measure participants' and targets' true traits. However, it is worth considering the potential implications of a self-report approach particularly in relation to our measurement of participant's self-perception accuracy.

There are four possible outcomes of comparing participants' HEXACO factor scores with their perception of their own HEXACO traits. First, participants may be genuinely accurate in their self-perception: their reported self-perception may be consistent with their HEXACO factor scores, and these factor scores may be genuinely reflective of their true traits. In this case, there is no doubt surrounding the accuracy of their self-perception. Second, participants may report traits that are inconsistent with their HEXACO factor scores, when these factor scores are indeed reflective of their true traits. These participants clearly have mis-located themselves in Mind-space and are likely to mis-locate a similar target. They should recognise the target's behaviour as like their own and locate them in the location they erroneously represent themselves as occupying. Given the well-documented reliability and predictive validity of the HEXACO-PI-R, we consider these first two outcomes to be the most likely in the present study, and as such our interpretation of our observed effects should be considered primarily in terms of these two possibilities, but two others are logically possible and thus warrant discussion.

A third, perhaps less likely outcome, is that a participant's perception of their own traits may be inconsistent with their HEXACO factor scores and that this inconsistency may arise because their HEXACO factor scores are incorrect, due to the participant having an impairment in predicting or remembering their own behaviour (and thus completing the HEXACO questionnaire incorrectly). Consequently, their self-perceived traits may be more indicative of their true traits than their responses to the HEXACO-60. In this case, our self-perception measure would indicate that participants have poor self-perception. Specifically, these participants would have poor self-perception in terms of their ability to predict their own behaviour, but not in their ability to locate themselves in Mind-space. These participants would also be expected to show a reduced similarity effect, but through a different mechanism to that described above. In this case, participants may observe the behaviour of a similar other and fail to recognise that the target's behaviour matches their own likely response in the same situation. The target truly occupies a similar region of Mind-space to the participant's (accurate) self-perception, but the participant, failing to recognise their similarity, would locate them elsewhere and thus be inaccurate in their trait inference. As such, consideration of this third possible outcome suggests that any disparity between self-perception and HEXACO factor scores should be associated with a reduced similarity effect, as observed in the present study.

The final possible outcome of comparing participants' perceived traits with their HEXACO factor scores is that these values are consistent even in the presence of inaccurate self-perception and behavioural prediction. Specifically, participants may be poor at predicting their own behaviour and locate themselves in Mind-space on the basis of these inaccurate predictions. Despite participants having poor self-perception, this pattern of responses would not be associated with a reduced similarity effect. If participants mis-represent their traits and mis-predict their behaviour in a consistent manner, they should show a similarity effect for targets who have traits and show behaviours that are similar to their self-perception, even if that perception is erroneous. If they observe a target who behaves in the manner that they expect that they themselves would, they should locate this target close to where they locate themselves in Mind-space. Given the location and the behaviour are consistent, even if not accurate in regard to the participant themselves, the resultant trait inference should be accurate for the traits of the target.

Therefore, the present results are to be expected under the Mind-space framework even if our measurements of the accuracy of participants' self-perception cannot fully differentiate between the four possible patterns of responses. We cannot confidently claim that consistent responses across the HEXACO-60 and reported self-perception on trait dimensions are definitely reflective of truly accurate self-perception, or that disparate responses necessarily reflect accurate behavioural predictions paired with inaccurate self-location in Mind-space. However, empirical investigations of the HEXACO-PI-R suggest that this is the most likely case. Regardless, further investigation is required to distinguish between these possibilities, most notably because differences in both self-location in Mind-space and in behavioural prediction in self-report personality inventories are possible sources of individual differences in understanding the traits and mental states of oneself and of others. Such investigation would likely need to test participants' predictions about their own behaviour against true behaviours that could be observed in an experimental setting, or through some form of experience sampling.

Results from our analyses regarding participants' estimates of population median traits and the interaction between participant-target similarity and participant self-perception accuracy support the idea, in accordance with the Mind-space framework (Conway et al., 2019), that the structure of a mentaliser's Mind-space and their ability to locate others within that space are experience-dependent. In the present study, we tested this using the HEXACO six personality dimensions (Ashton & Lee, 2007; Ashton et al., 2014; Lee & Ashton, 2008). It should be noted that the Mind-space framework does not make specific predictions regarding which (or how many) trait dimensions constitute Mind-space. Instead, the theory suggests that the dimensions which comprise a mentaliser's Mind-space are those which have been learned (by that mentaliser) to enable minds to be individuated (perhaps in part for the purposes of allowing accurate metal state inference).

The factor-analytic methods used to identify the HEXACO six personality dimensions necessarily imply that these dimensions constitute an effective method of representing a wide array of possible trait descriptors (as taken from lexical studies, (Ashton & Lee, 2007)) and/or typical behaviours (as obtained from questionnaire measures, (Ashton & Lee, 2009)) in a reduced dimensional form. As such, these trait dimensions provide a large amount of information for the prediction of mental states in a condensed form, and it is thus expected that these dimensions should form at least part of most individuals' Mind-spaces. It is for this reason that these dimensions were used in the present study. Other dimensions, including cognitive dimensions (e.g., IQ, working memory), may be represented in Mind-space, and individuals may use a larger number of more specific trait dimensions (e.g., those often considered as facets of factor-level dimensions (Ashton & Lee, 2007)) to gain additional information diagnostic of mental states.

4.3. Predictors of confidence in trait inference

Having determined that similarity and self-perception accuracy are associated with the accuracy of trait inferences, we investigated whether similarity and self-perception accuracy might play a role in the construction of confidence in trait inferences. Given that similarity is associated with more accurate trait inferences, we hypothesised that participants might have learned to use similarity as a cue from which they could determine the likelihood that a given inference was accurate, and thus their confidence in that inference. The tendency to be generally overconfident, rather than underconfident, in one's performance is well documented (Baranski & Petrusic, 1995; Brenner, Koehler, Liberman, & Tversky, 1996; Dunning, Griffin, Milojkovic, & Ross, 1990; Hoffrage, 2017; Moore & Schatz, 2017). Therefore, we theorised that, if similarity is used as a cue for confidence, participants might be more confident in their inferences than is warranted by their accuracy when the target is more similar to them. We would therefore expect the relationship between confidence and error to be less negative (i.e., for confidence to reduce less as error increases) when the target is more similar to the participant.

In addition, given that we found, as predicted, that self-perception accuracy influences the extent to which participants gain the potential benefit of similarity (i.e., the extent to which their inferences are more accurate for those more similar to them), we expected that self-perception accuracy would also modulate the effect of similarity on the relationship between trait inference error and confidence. The present study indicates that this is indeed the case, as we found a three-way interaction between trait inference error, participant-target trait difference, and participant self-perception error when predicting confidence in trait inferences.

To illustrate, consider a mentaliser with an erroneous perception of their location on a trait dimension (e.g., extraversion). This mentaliser may still be more confident in their trait inference when locating a similar other in Mind-space but (according to the findings outlined earlier) would also be likely to make a more erroneous trait inference than an individual with more accurate self-perception. In this case, we would expect the overconfidence observed in trait inferences about similar others (i.e., the presence of a less negative relationship between confidence and error when inferences are made about targets more similar to the mentaliser) to be further amplified when the mentaliser has a more erroneous perception of their own location on the trait dimension in question. In other words, the increase in confidence arising from similarity between the target and the mentaliser would be (further) misplaced, because a mentaliser with poorer self-perception gains less of a similarity benefit in the accuracy of their inference.

One might instead have predicted, however, that a mentaliser with inaccurate self-perception of their traits may not have learned to use similarity as a cue to confidence. This would be expected if their similarity to the targets they encounter in everyday life does not predict, in general, the accuracy with which they can infer that target’s traits, mental states, or behaviour. However, given that the similarity benefit is observed, albeit to a lesser degree, when self-perception accuracy is poor, and the fact that most individuals likely have relatively accurate self-perception in some, even if not all, personality dimensions, the Mind-space framework would predict that most people would learn to use similarity as a cue to confidence. It remains the case, though, that the extent to which similarity influences confidence might be determined by the extent to which, in each mentaliser’s personal experience, it is diagnostic of accuracy. Exploring individual differences in the construction of confidence in trait inference, including the influence of similarity, might therefore be a fruitful avenue for future work.

It might be considered somewhat surprising that the data plotted in Figure 8 indicate a positive correlation between confidence and error in trait inferences regarding similar others. The AUROC2 measure demonstrated that participants' confidence reports discriminate correct from incorrect answers at an above chance rate. Given this, it is perhaps counterintuitive that they appear to be more confident in more erroneous inferences. This pattern of results might be explained by overconfidence bias, a long-studied effect in which people tend to be more confident in their performance than is justified by the performance itself (Baranski & Petrusic, 1995; Brenner et al., 1996; Dunning et al., 1990; Hoffrage, 2017; Moore & Schatz, 2017).

Overconfidence is known to be greater when participant estimates are further from population base levels (Dunning et al., 1990). Whilst Dunning et al. (1990) identified this effect as a result of a reduction in confidence smaller than the reduction in error as estimates diverge from base levels, in the case of our task, participants appeared to be more confident when making more extreme trait inferences (i.e., when they judged the target to be well below or well above the population median on a given trait). Indeed, a supplementary analysis indicated that confidence increased as the difference between participants' estimates of targets' traits and the population median for that trait increased ($B = 0.31$, $SE = 0.01$, $t\,(9799.68) = 36.80$, $p < .001$). This effect appears to occur within participants, as a similar increase in confidence was observed as the difference between a participant's trait estimate on a given trial and the mean estimate made by that participant across all trials increased ($B = 0.26$, $SE = 0.01$, $t\,(9745.91) = 29.22$, $p < .001$). One possible explanation for this effect is that cues that a target is highly extraverted or highly introverted, for example, might be more salient than behaviours indicating 'average' levels of extraversion. Full details of these supplementary analyses are given in the Supplementary Materials (Section S.6.).

Statistically, the further one's estimate is from the population median, the more inaccurate that estimate is likely to be. Therefore, given that participants are more confident in more extreme inferences, we would expect to see a positive relationship between confidence and accuracy, because more extreme inferences are likely, on average, to hold a higher degree of error. In contrast, the AUROC2 measure should not be affected by overconfidence in extreme judgements, because it quantified whether participants were more likely to be confident in inferences which correctly identified the target as above or below the population median. It seems, then, that although participants had sufficient insight into the accuracy of their judgements that their confidence discriminated between trials in which they were correct or incorrect about the direction of the target's difference from the population median, they were ultimately overconfident. This overconfidence was heightened when the participants perceived targets to have more extreme levels of a trait, and when the target was more similar to the participant.

Although the AUROC2 measure should not be affected by heightened overconfidence in more extreme trait judgements, these results do indicate that metacognitive sensitivity is likely to be affected by characteristics of the target and the participant. If, as these results indicate, participants are using similarity as a cue to confidence, with different levels of success according to the accuracy of their self-perception, then there are several factors which would be expected to influence their measured metacognitive ability. We have explored two of these in the present work: the participant's perception of their own traits relative to the true values of those traits (i.e., their self-perception accuracy); and the traits of the targets included in the stimuli relative to the participant (i.e., participant-target similarity). A third factor, the traits of the targets included in the stimuli relative to the participant's perception of their own traits, may also be important. It is possible that mentalisers with poor self-perception might accurately locate others in the location they (wrongly) perceive themselves to occupy – meaning that they may show a similarity benefit not for those who are truly similar to them, but those that they believe to be similar to them.

Given this, it is plausible, perhaps even to be expected, that participants would have different measured levels of metacognitive sensitivity with different sets of stimuli. This might go some way to explaining why we did not observe an association between participant AUROC2 score and participant average mental state inference error in our linear regression analysis. One possibility is that the targets a participant viewed in the four videos of the Interview Task may not be representative of the broader corpus of video stimuli used in the metacognition task. Indeed, as each participant saw four videos randomly selected from a broader set, we would expect the Interview Task targets to be representative of the full video corpus *across* participants, but not every participant would be expected to observe a representative set. The AUROC2 measure, then, might capture both general metacognitive sensitivity in the domain of trait inference, *and* target-specific metacognitive sensitivity for the set of targets observed. Future research using multiple distinct stimulus sets would help to disentangle these two aspects of metacognitive sensitivity in trait inference.

However, our mechanistic linear mixed model analysis indicated that those with greater measured metacognitive ability in the metacognition task did report confidence levels that were more in line with their trait inference accuracy and weight their trait inferences accordingly. This analysis accounted for features of the stimuli in a way our multiple linear regression could not. Specifically, conducting a more sensitive, trial-by-trial analysis including trait inference error and confidence (alongside random intercepts for participant, target and trial) means that our model was able to account for target-specific differences in each participants' trait inference error and confidence. The variance explained by AUROC2 in interaction with trait inference error and confidence (i.e., the predicted three-way interaction) therefore indicates that, when target-specific differences are accounted for, greater metacognitive sensitivity does support more optimal weighting of trait inferences in the process of mental state inference.

It seems, therefore, that metacognition does play an important role in the weighting of trait information in mental state inference, but that there may not be one unitary 'metacognitive ability' within the trait inference domain. This being the case, one must consider what is underlying the association between metacognitive ability (measured in domains less clearly related to ToM ability, such as perception or memory) and ToM ability in those studies in which it is observed (K. L. Carpenter et al., 2019; Nicholson et al., 2020; van der Plas et al., 2021; D. M. Williams et al., 2018). One possibility, as outlined in the Introduction, is that the association between metacognitive ability and ToM ability found in these studies may have resulted from some third factor which influences measurements of both abilities, such as average confidence or perceptual or sensory differences.

However, as previously mentioned, it appears that metacognition is likely to consist of both domain-specific and domain-general components (J. Carpenter et al., 2019; Fitzgerald et al., 2017; Fleming et al., 2014; Morales et al., 2018; Rouault et al., 2018). It is possible that previous work has captured a domain-general component which may be associated with mental state inference accuracy through the optimisation of the weighting of trait inferences that we have described alongside other routes. For example, metacognitive ability may also influence the use of inferred situational information; the use of perceptual cues, such as facial expression or vocal intonation; or the way in which one learns from experience regarding the relationships between traits, situations, and mental states.

The metacognition task used in the present study might not isolate this domain-general component in the same manner as tasks in other domains. Whereas domain-specific or stimulus-specific differences in ability in perceptual or memory domains might be more limited (as stimuli are able to be standardised in a way that is not viable in the present context) or might appear as noise or measurement error when predicting ToM ability, these differences in the trait inference domain are very much relevant to the accuracy of mental state inference in the Interview Task. As such, it seems that domain-general metacognitive ability may be overshadowed by domain- and stimulus- specific abilities in our metacognition task, and it is only by accounting for these that the effect of metacognitive sensitivity on mental state inference accuracy becomes clear.

There remain, however, important questions to confront regarding the extent to which measurements of metacognitive ability are best considered as representative of individual differences in a stable ability. Whilst it is possible that other studies of the relationship between metacognition and ToM have captured, in their measures, a domain-general metacognitive ability that is associated with ToM ability, our results make clear that any such general ability is only part of the picture. Moreover, recent work by Rahnev (2023) shows that, across all commonly-used metacognition measures, test-retest reliability is low despite split-half reliability being relatively high. Even in these existing measures, then, it seems that the metacognitive ability being captured is not a unitary, stable ability, but one that may be highly influenced by state effects (e.g., participants' level of alertness on the day of testing) or other temporal effects (e.g., experience or practice effects). Considering this, alongside the evidence that we present here, it seems that to understand the role of metacognition in ToM (and indeed in cognition more broadly), the field might benefit from considering metacognition as a process, the effectiveness of which can vary for many reasons, more so than as a source of stable individual differences in ability.

4.4. Conclusions

The present study sought to investigate the mechanisms underlying ToM inferences, specifically examining the role of metacognition, trait inference, and possible predictors of trait inference ability and confidence in trait inferences. The conclusions of this study are illustrated in Figure 9.

First, we replicated the finding that more accurate trait inferences are associated with more accurate mental state inferences. Then, we found that metacognitive ability facilitates more accurate mental state inference. Specifically, we found evidence that mentalisers weight their trait inferences according to their confidence, relying more heavily on trait inferences in which they are more confident. Whilst we did not find a simple association between metacognitive ability and ToM ability, results indicated that better metacognitive ability facilitates more optimal weighting of trait inferences. This effect emerges because those with better metacognitive sensitivity tend, when characteristics of the target are accounted for, to be more confident in accurate inferences and less confident in inaccurate inferences.

We also examined factors which were thought to influence the accuracy of trait inferences themselves. We found that similarity and the accuracy of self-perception interact in predicting the accuracy with which participants are able to locate targets in Mind-space, such that participants with accurate self-perception showed a greater reduction in trait inference error when locating a more similar target in Mind-space than participants with less accurate self-perception. Furthermore, results indicated that the accuracy of participants' perceptions of the 'average' mind are also associated with the accuracy of trait inference.

In addition, we found that the similarity between the target and the mentaliser influences not only the accuracy with which the mentaliser can locate the target in Mind-space, but also their confidence in this judgement, such that participants were more likely to be overconfident in a judgement when they were more similar to the target. We found that self-perception accuracy impacts the extent to which this influence is beneficial. Through modulating the extent of the similarity benefit in trait inference accuracy, the accuracy of self-perception also, in turn, affects the degree to which the mentaliser's level of confidence reflects the accuracy of their judgement.

The results of this study are in accordance with the Mind-space framework (Conway et al., 2019),the core tenet of which is that mentalisers' perceptions of targets' traits are used in inferring targets' mental states. Furthermore, the associations between similarity, self-perception accuracy and the understanding of the average mind with trait inference accuracy and confidence provide support for another central idea of the Mind-space theory: that learning from social experience shapes the structure of Mind-space itself, the ability to locate targets within that space, and the way in which Mind-space location is used to infer mental states. The present study serves to highlight how, as a result of this experience-dependence, characteristics of the target and the mentaliser play important roles in several aspects of mental state inference, including the accuracy of the information on which mental state inferences are based and the way in which that information is used.

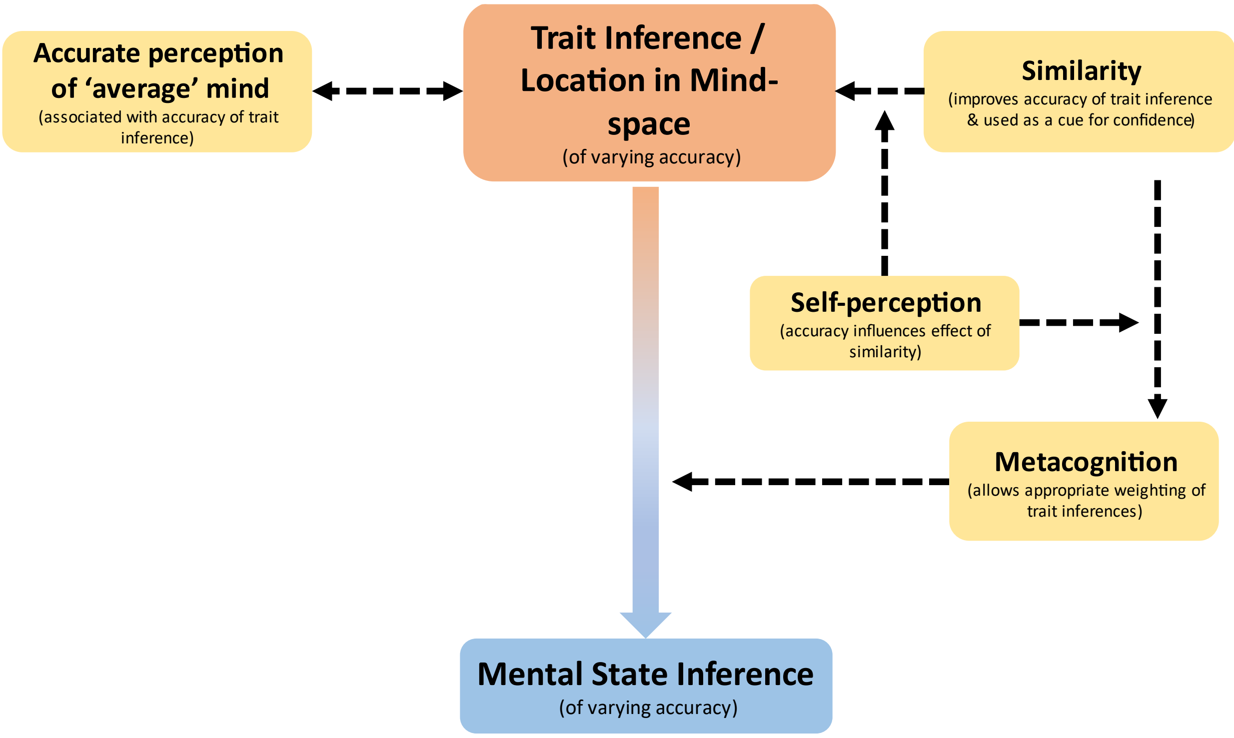

Figure 9. A schematic of processes thought to be involved in accurate mental state inference based on the present study.

Acknowledgements

This publication was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. E.L.L. was supported by the Economic and Social Research Council (ES/P000649/1). S.M.F. is funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [selected as ERC Consolidator, grant number 101043666]. S.M.F. is a CIFAR Fellow in the Brain, Mind & Consciousness Program. The funding sources that supported this article were not involved in study design; in the collection, analysis or interpretation of data; in the writing of the report; or in the decision to submit the article for publication. The authors would like to thank Lauren Charters for support obtaining video stimuli for this study.

Data Statement

The datasets generated and analysed during the current study are not publicly available due to participant restrictions on data sharing, but shareable data are available from the corresponding author on reasonable request.

References

Abell, F., Happé, F., & 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. doi:https://doi.org/10.1016/S0885-2014(00)00014-9

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388-407. doi:10.3758/s13428-019-01237-x

Ashton, M. C., & Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. *Personality and Social Psychology Review, 11*(2), 150-166. doi:10.1177/1088868306294907

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A Short Measure of the Major Dimensions of Personality. *Journal of Personality Assessment, 91*(4), 340-345. doi:10.1080/00223890902935878

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review, 18*(2), 139-152.

Ashwin, E., Ashwin, C., Rhydderch, D., Howells, J., & Baron-Cohen, S. (2009). Eagle-Eyed Visual Acuity: An Experimental Investigation of Enhanced Perception in Autism. *Biological Psychiatry, 65*(1), 17-21. doi:https://doi.org/10.1016/j.biopsych.2008.06.012

Ashwood, K. L., Gillan, N., Horder, J., Hayward, H., Woodhouse, E., McEwen, F. S., . . . Murphy, D. G. (2016). Predicting the diagnosis of autism in adults using the Autism-Spectrum Quotient (AQ) questionnaire. *Psychological medicine, 46*(12), 2595-2604. doi:10.1017/S0033291716001082

Bang, D., Moran, R., Daw, N. D., & Fleming, S. M. (2022). Neurocomputational mechanisms of confidence in self and others. *Nature Communications, 13*(1), 4238. doi:10.1038/s41467-022-31674-w

Baranski, J. V., & Petrusic, W. M. (1995). On the Calibration of Knowledge and Perception. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale, 49*(3), 397-407. doi:10.1037/1196-1961.49.3.397

Baron-Cohen, S. (1990). Autism: A Specific Cognitive Disorder of "Mind-Blindness". *International Review of Psychiatry, 2*(1), 81-90. doi:10.3109/09540269009028274

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind” ? *Cognition, 21*(1), 37-46. doi:https://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J Child Psychol Psychiatry, 42*(2), 241-251.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5-17. doi:10.1023/A:1005653411471

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 10.1016/j.jml.2012.1011.1001. doi:10.1016/j.jml.2012.11.001

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology [Cambridge University Press]. Retrieved

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bird, G., & Cook, R. (2013). Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism. *Transl Psychiatry, 3*(7), e285. doi:10.1038/tp.2013.61

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in Probability and Frequency Judgments: A Critical Examination. *Organizational Behavior and Human Decision Processes, 65*(3), 212-219. doi:https://doi.org/10.1006/obhd.1996.0021

Brüne, M. (2005). "Theory of Mind" in Schizophrenia: A Review of the Literature. *Schizophrenia Bulletin, 31*(1), 21-42. doi:10.1093/schbul/sbi002

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research, 33*(2), 261-304.

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General, 148*(1), 51-64. doi:10.1037/xge0000505

Carpenter, K. L., Williams, D. M., & Nicholson, T. (2019). 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. doi:10.1007/s10803-019-04118-6

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and brain sciences, 32*(2), 121-138. doi:10.1017/S0140525X09000545

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*: OUP Oxford.

Carruthers, P., & Smith, P. K. (1996). *Theories of theories of mind*: Cambridge university press.

Christensen, R. H. B. (2023). ordinal - Regression Models for Ordinal Data (Version R package version 2023.12-4). Retrieved from https://cran.r-project.org/package=ordinal

Clarke, F., Birdsall, T., & Tanner Jr, W. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America, 31*(5), 629-630.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*: Routledge.

Cohen, J. (1992). A power primer. *Psychol Bull, 112*(1), 155-159. doi:10.1037//0033-2909.112.1.155

Constantino, J. N., & Todd, R. D. (2003). Autistic Traits in the General Population: A Twin Study. *Archives of General Psychiatry, 60*(5), 524-530. doi:10.1001/archpsyc.60.5.524

Conway, J. R., Catmur, C., & Bird, G. (2019). Understanding individual differences in theory of mind via representation of minds, not mental states. *Psychon Bull Rev, 26*(3), 798-812. doi:10.3758/s13423-018-1559-x

Conway, J. R., Coll, M. P., Cuve, H. C., Koletsi, S., Bronitt, N., Catmur, C., & Bird, G. (2020). Understanding how minds vary relates to skill in inferring mental states, personality, and intelligence. *J Exp Psychol Gen, 149*(6), 1032-1047. doi:10.1037/xge0000704

de Vries, R. E., Pronk, J., Olthof, T., & Goossens, F. A. (2020). Getting along And/Or Getting Ahead: Differential Hexaco Personality Correlates of Likeability and Popularity among Adolescents. *European Journal of Personality, 34*(2), 245-261. doi:10.1002/per.2243

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction [American Psychological Association doi:10.1037/0022-3514.58.4.568]. Retrieved

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Brief Bioinform, 21*(2), 553-565. doi:10.1093/bib/bbz016

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., . . . Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *J Autism Dev Disord, 36*(5), 623-636. doi:10.1007/s10803-006-0107-0

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Fitzgerald, L. M., Arvaneh, M., & Dockree, P. M. (2017). Domain-specific and domain-general processes underlying metacognitive judgments. *Consciousness and Cognition, 49*, 264-277. doi:https://doi.org/10.1016/j.concog.2017.01.011

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological review, 124*(1), 91.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*. doi:10.3389/fnhum.2014.00443

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain, 137*(10), 2811-2822. doi:10.1093/brain/awu221

Frith, C. D., & Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychological medicine, 26*(3), 521-530. doi:10.1017/S0033291700035601

Galvin, S. J., Podd, J. V., Drga, V., & 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.

Georghiades, P. (2004). From the general to the situated: three decades of metacognition. *International Journal of Science Education, 26*(3), 365-383. doi:10.1080/0950069032000119401

Gertler, B. (2010). *Self-knowledge*: Routledge.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*: Oxford University Press on Demand.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and brain sciences, 16*(1), 1-14.

Grainger, C., Williams, D. M., & Lind, S. E. (2014). Metacognition, metamemory, and mindreading in high-functioning adults with autism spectrum disorder. *Journal of abnormal psychology, 123*(3), 650-659. doi:10.1037/a0036531

Grainger, C., Williams, D. M., & Lind, S. E. (2016). Metacognitive monitoring and control processes in children with autism spectrum disorder: Diminished judgement of confidence accuracy. *Conscious Cogn, 42*, 65-74. doi:10.1016/j.concog.2016.03.003

Griffin, J. W., Bauer, R., & Gavett, B. E. (2022). The Episodic Memory Profile in Autism Spectrum Disorder: A Bayesian Meta-Analysis. *Neuropsychology Review, 32*(2), 316-351. doi:10.1007/s11065-021-09493-5

Gumley, A. (2011). Metacognition, affect regulation and symptom expression: A transdiagnostic perspective. *Psychiatry research, 190*(1), 72-78.

Happé, F. (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. doi:10.1007/BF02172093

Happé, F. (2003). Theory of mind and the self. *Annals of the New York Academy of Sciences, 1001*(1), 134-144.

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making, 13*(4), 356-371.

Hill, E., Berthoz, S., & Frith, U. (2004). Brief report: cognitive processing of own emotions in individuals with autistic spectrum disorder and in their relatives. *J Autism Dev Disord, 34*(2), 229-235. doi:10.1023/b:jadd.0000022613.41399.14

Hoekstra, R. A., Vinkhuyzen, A. A. E., Wheelwright, S., Bartels, M., Boomsma, D. I., Baron-Cohen, S., . . . van der Sluis, S. (2011). The construction and validation of an abridged version of the autism-spectrum quotient (AQ-Short). *Journal of Autism and Developmental Disorders, 41*(5), 589-596. doi:10.1007/s10803-010-1073-0

Hoffrage, U. (2017). Overconfidence. In *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory, 2nd ed.* (pp. 291-314). New York, NY, US: Routledge/Taylor & Francis Group.

Johnstone, A., Friston, K., Rees, G., & Lawson, R. P. (2022). Metacognitive and noradrenergic differences in autistic adults.

Jussila, K., Junttila, M., Kielinen, M., Ebeling, H., Joskitt, L., Moilanen, I., & Mattila, M. L. (2020). Sensory Abnormality and Quantitative Autism Traits in Children With and Without Autism

Spectrum Disorder in an Epidemiological Population. *Journal of Autism and Developmental Disorders, 50*(1), 180-188. doi:10.1007/s10803-019-04237-0

Körding, K. (2007). Decision theory: what "should" the nervous system do? *Science, 318*(5850), 606-610. doi:10.1126/science.1142998

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software; Vol 1, Issue 13 (2017)*. doi:10.18637/jss.v082.i13

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate behavioral research, 39*(2), 329-358.

Lee, K., & Ashton, M. C. (2008). The HEXACO personality factors in the indigenous personality lexicons of English and 11 other languages. *Journal of personality, 76*(5), 1001-1054.

Lee, K., & Ashton, M. C. (2013). Prediction of self- and observer report scores on HEXACO-60 and NEO-FFI scales. *Journal of Research in Personality, 47*(5), 668-675. doi:https://doi.org/10.1016/j.jrp.2013.06.002

Lee, K., & Ashton, M. C. (2018). Psychometric Properties of the HEXACO-100. *Assessment, 25*(5), 543-556. doi:10.1177/1073191116659134

Lee, K., Ashton, M. C., Griep, Y., & Edmonds, M. (2018). Personality, Religion, and Politics: An Investigation in 33 Countries. *European Journal of Personality, 32*(2), 100-115. doi:10.1002/per.2142

Leslie, A. M., & Frith, U. (1987). Metarepresentation and autism: how not to lose one's marbles. *Cognition, 27*(3), 291-294. doi:10.1016/s0010-0277(87)80014-8

Long, E. L., Cuve, H. C., Conway, J. R., Catmur, C., & Bird, G. (2022). Novel theory of mind task demonstrates representation of minds in mental state inference. *Scientific Reports, 12*(1), 21133. doi:10.1038/s41598-022-25490-x

Maehara, Y., & Umeda, S. (2013). Reasoning bias for the recall of one's own beliefs in a Smarties task for adults. *Japanese Psychological Research, 55*(3), 292-301. doi:https://doi.org/10.1111/jpr.12009

McCrae, R. R., & Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective*: Guilford Press.

McMahon, C. M., Henderson, H. A., Newell, L., Jaime, M., & Mundy, P. (2016). Metacognitive Awareness of Facial Affect in Higher-Functioning Children and Adolescents with Autism Spectrum Disorder. *J Autism Dev Disord, 46*(3), 882-898. doi:10.1007/s10803-015-2630-3

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass, 11*(8), e12331. doi:https://doi.org/10.1111/spc3.12331

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. *The Journal of Neuroscience, 38*(14), 3534-3546. doi:10.1523/jneurosci.2360-17.2018

Moshagen, M., Thielmann, I., Hilbig, B., & Zettler, I. (2019). Meta-Analytic Investigations of the HEXACO Personality Inventory(-Revised): Reliability Generalization, Self–Observer Agreement, Intercorrelations, and Relations to Demographic Variables. *Zeitschrift für Psychologie, 227*, 186-194. doi:10.1027/2151-2604/a000377

Mueller, R., Utz, S., Carbon, C.-C., & Strobach, T. (2020). Face Adaptation and Face Priming as Tools for Getting Insights Into the Quality of Face Space. *Frontiers in Psychology, 11*. doi:10.3389/fpsyg.2020.00166

Nichols, S., & Stich, S. P. (2003). *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds*: Clarendon Press/Oxford University Press.

Nicholson, T., Williams, D. M., Lind, S. E., Grainger, C., & Carruthers, P. (2020). Linking metacognition and mindreading: Evidence from autism and dual-task investigations. *Journal of Experimental Psychology: General*. doi:10.1037/xge0000878

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of abnormal psychology, 125*(6), 818-823. doi:10.1037/abn0000182

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences, 1*(4), 515-526. doi:10.1017/S0140525X00076512

Proust, J. (2007). Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? *Synthese, 159*(2), 271-295. doi:10.1007/s11229-007-9208-3

R Core Team. (2020). R: A language and environment for statistical computing. . Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rahnev, D. (2023). Measuring metacognition: A comprehensive assessment of current methods.

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience, 1*, e17. doi:10.1017/pen.2018.16

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism, 6*(1), 2. doi:10.1186/2040-2392-6-2

Sifneos, P. E. (1973). The prevalence of "alexithymic" characteristics in psychosomatic patients. *Psychotherapy and Psychosomatics, 22*(2-6), 255-262. doi:10.1159/000286529

Sizoo, B. B., Horwitz, E. H., Teunisse, J. P., Kan, C. C., Vissers, C., Forceville, E. J. M., . . . Geurts, H. M. (2015). Predictive validity of self-report questionnaires in the assessment of autism spectrum disorders in adults. *Autism, 19*(7), 842-849. doi:10.1177/1362361315589869

Southwick, J. S., Bigler, E. D., Froehlich, A., DuBray, M. B., Alexander, A. L., Lange, N., & Lainhart, J. E. (2011). Memory functioning in children and adolescents with autism. *Neuropsychology, 25*(6), 702-710. doi:10.1037/a0024935

Soutter, A. R. B., Bates, T. C., & Mõttus, R. (2020). Big Five and HEXACO Personality Traits, Proenvironmental Attitudes, and Behaviors: A Meta-Analysis. *Perspectives on Psychological Science, 15*(4), 913-941. doi:10.1177/1745691620903019

Takarae, Y., Sablich, S. R., White, S. P., & Sweeney, J. A. (2016). Neurophysiological hyperresponsivity to sensory input in autism spectrum disorders. *Journal of Neurodevelopmental Disorders, 8*(1), 29. doi:10.1186/s11689-016-9162-9

Taylor, G. J., Bagby, R. M., & Parker, J. D. (2003). The 20-Item Toronto Alexithymia Scale: IV. Reliability and factorial validity in different languages and cultures. *Journal of psychosomatic research, 55*(3), 277-283.

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin, 146*(1), 30.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-Space: A Unifying Concept in Face Recognition Research. *Quarterly Journal of Experimental Psychology, 69*(10), 1996-2019. doi:10.1080/17470218.2014.990392

van der Plas, E., Mason, D., Livingston, L. A., Craigie, J., Happé, F., & Fleming, S. M. (2021). Computations of confidence are modulated by mentalizing ability. doi:doi.org/10.31234/osf.io/c4pzj

Washburn, D., Wilson, G., Roes, M., Rnic, K., & Harkness, K. L. (2016). Theory of mind in social anxiety disorder, depression, and comorbid conditions. *Journal of Anxiety Disorders, 37*, 71-77. doi:https://doi.org/10.1016/j.janxdis.2015.11.004

Wechsler, D. (2011). Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II) San Antonio. *TX: Pearson.[Google Scholar]*.

Wilkinson, D. A., Best, C. A., Minshew, N. J., & Strauss, M. S. (2010). Memory awareness for faces in individuals with autism. *J Autism Dev Disord, 40*(11), 1371-1377. doi:10.1007/s10803-010-0995-x

Williams, D. L., Goldstein, G., & Minshew, N. J. (2006). The profile of memory function in children with autism. *Neuropsychology, 20*(1), 21-29. doi:10.1037/0894-4105.20.1.21

Williams, D. M., Bergström, Z., & 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. doi:10.1177/1362361316680178

Williams, Z. J., Suzman, E., Bordman, S. L., Markfeld, J. E., Kaiser, S. M., Dunham, K. A., . . . Woynaroski, T. G. (2022). Characterizing Interoceptive Differences in Autism: A Systematic Review and Meta-analysis of Case–control Studies. *Journal of Autism and Developmental Disorders*. doi:10.1007/s10803-022-05656-2

Wilson, T. D. (2004). *Strangers to ourselves*: Harvard University Press.

Wojcik, D. Z., Allen, R. J., Brown, C., & Souchay, C. (2011). Memory for actions in autism spectrum disorder. *Memory, 19*(6), 549-558. doi:10.1080/09658211.2011.590506

Wojcik, D. Z., Moulin, C. J., & Souchay, C. (2013). Metamemory in children with autism: Exploring “feeling-of-knowing” in episodic and semantic memory. *Neuropsychology, 27*(1), 19.

Wuttke, S. J., & Schweinberger, S. R. (2019). The P200 predominantly reflects distance-to-norm in face space whereas the N250 reflects activation of identity-specific representations of known faces. *Biol Psychol, 140*, 86-95. doi:10.1016/j.biopsycho.2018.11.011

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1310-1321.

Zalla, T., Miele, D., Leboyer, M., & Metcalfe, J. (2015). Metacognition of agency and theory of mind in adults with high functioning autism. *Consciousness and Cognition, 31*, 126-138. doi:https://doi.org/10.1016/j.concog.2014.11.001