Consciousness and Cognition 70 (2019) 11–24

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Relationships between implicit and explicit uncertainty monitoring and mindreading: Evidence from autism spectrum disorder

Toby Nicholson[a], David M. Williams[a,*], Catherine Grainger[b], Sophie E. Lind[c], Peter Carruthers[d]

[a] School of Psychology, University of Kent, UK
[b] School of Psychology, University of Stirling, UK
[c] Department of Psychology, City, University of London, UK
[d] Department of Philosophy, University of Maryland, USA

ARTICLE INFO

Keywords:
Autism spectrum disorder
Metacognition
Mindreading
Uncertainty judgment
Uncertainty monitoring

ABSTRACT

We examined performance on implicit (non-verbal) and explicit (verbal) uncertainty-monitoring tasks among neurotypical participants and participants with autism, while also testing mind-reading abilities in both groups. We found that: (i) performance of autistic participants was unimpaired on the implicit uncertainty-monitoring task, while being significantly impaired on the explicit task; (ii) performance on the explicit task was correlated with performance on mindreading tasks in both groups, whereas performance on the implicit uncertainty-monitoring task was not; and (iii) performance on implicit and explicit uncertainty-monitoring tasks was not correlated. The results support the view that (a) explicit uncertainty-monitoring draws on the same cognitive faculty as mindreading whereas (b) implicit uncertainty-monitoring only test first-order decision making. These findings support the theory that metacognition and mind-reading are underpinned by the same meta-representational faculty/resources, and that the implicit uncertainty-monitoring tasks that are frequently used with non-human animals fail to demonstrate the presence of metacognitive abilities.

1. Introduction

Humans are self-aware. By this we mean that they are capable not only of forming mental representations of themselves as living entities existing in a world that is separate from themselves, but also of forming representations of their own mental representations (i.e., meta-representations; Pylyshyn, 1978). In short, humans are meta*cognitive* creatures: they entertain thoughts about their own thoughts.[1]

* Corresponding author at: School of Psychology, Keynes College, University of Kent, Canterbury CT2 7NP, UK.
E-mail address: D.M.Williams@kent.ac.uk (D.M. Williams).

[1] Metacognition has always been defined by cognitive and developmental psychologists as “thought about thought” or “thinking about thinking” (Flavell, 1979; Nelson & Narens, 1990; Dunlosky & Metcalfe, 2009). More recently, some comparative psychologists have weakened this definition in an attempt to allow for forms of *implicit*, nonconceptual, self-awareness in nonhuman animals (Couchman et al., 2012; Smith et al., 2014). We will return to this point below. In what follows, by an implicit *task* we mean a purely-behavioral, non-verbal, task; whereas an explicit task involves a verbal response of some sort. An implicit mental *process*, in contrast, is one that doesn’t involve conceptual judgments; whereas an explicit process does. This leaves open the possibility that an implicit task might tap into an explicit (thought-involving) process, and also that explicit responses might depend, in part, on implicit/nonconceptual mental processes.

https://doi.org/10.1016/j.concog.2019.01.013
Received 21 December 2018; Received in revised form 11 January 2019; Accepted 18 January 2019

One important question concerns the relation between metacognition (awareness of one's own mental states) and mindreading or "theory of mind" (awareness of the mental states of others). This has been subject to intense debate among philosophers, and these debates have increasingly spilled into the scientific domain. There are broadly three positions one can take (and which have been taken) concerning the relation between metacognition and mindreading.

The first has a long tradition in philosophy, dating back at least to Descartes (1637, 1641). It is that we are aware of our own minds through introspection, and that we combine this self-awareness together with other forms of knowledge, and other kinds of ability, in order to know of the mental states of others. To have knowledge of others' mental states, for example, one might rely on generalizations about the links between mental states and behavior that have been learned from one's own case. Or one might use one's imaginative abilities to figure out what someone else is seeing from a different perspective. In its most recent (empirically-informed) incarnation, the view is that to engage in successful mindreading, we employ imaginative *simulations* of the perspectives of other people, attributing mental states to them by introspecting in ourselves the results of these simulations (Goldman, 2006). Inherent to this view is that we have a privileged, non-inferential access to our own mental states via introspection, whereas mindreading is based on use of metacognition via mental simulation. We will refer to this as the "self-awareness is prior" account.

A second possibility is that we are aware of our own mental states by introspection, but we employ a distinct mindreading faculty when reasoning about the mental states of other people (Nichols & Stich, 2003). On this view, the two capacities might have distinct evolutionary origins and paths, are likely to emerge separately in child development, and will be uncorrelated in task performance. We will call this the "two systems" view.

Then finally, it has been claimed that there is no special faculty of introspection. Rather, self-knowledge and other-knowledge depend on the operations of the same underlying faculty/process of meta-representation, albeit relying on partially distinct inputs. For example, when attributing mental states to the self, the meta-representational faculty has access to one's own inner speech and visual imagery, whereas the faculty only has access to observable behavior when attributing mental states to others. According to this view, this meta-representational faculty/process evolved in the first instance for other-directed mindreading, and this deployment of it is also the first to emerge in child development (Carruthers, 2009, 2011; Gopnik, 1993; Wellman, Cross, & Watson, 2001). We will refer to this as the "one system" view.

Many possible lines of evidence can be brought to bear on the debate among these three views. First, evidence could be sought from studies exploring the association between mindreading and metacognition. Both the self-awareness-is-prior and the one-system views predict that measures of metacognition and mindreading should be correlated, since each maintains that the two abilities share processing resources (introspective ability, on the former account; a meta-representation faculty, on the latter). The two-systems view, in contrast, predicts that such measures should *not* be correlated significantly.

Second, evidence could be sought from studies of metacognition in populations that have mindreading impairments. Arguably the "classic" disorder of mindreading is autism spectrum disorder (ASD).[2] ASD is a neurodevelopmental disorder diagnosed on the basis of behavioral impairments in social-communication and behavioral flexibility (repetitive and restricted behavior and interests) (American Psychiatric Association, 2013). At the cognitive level, it is well established that autistic people have difficulties with mindreading (e.g., Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998) and that these difficulties contribute to at least some of the social-communication difficulties that are diagnostic of the disorder (see Brunsdon & Happé, 2014). Mindreading task performance is reliably diminished among autistic people relative to the performance of age- and IQ-matched non-autistic individuals (with a large associated effect size of $d = 0.88$, according to meta-analysis; see Yirmiya et al., 2001) and is associated with reduced activation of a well-defined mindreading brain network (including the temporo-parietal junction, medial prefrontal cortex, precuneus and inferior frontal gyrus; e.g., Castelli, Frith, Happé, & Frith, 2002; see Schurz, Radua, Aichhorn, Richlan, & Perner, 2014, for review and meta-analysis).

Given that individuals with ASD have clear difficulties meta-representing others' mental states, the study of metacognition in this disorder has the potential to inform theory, as well as our understanding of ASD itself. *Only* the one-system view predicts that autistic people should be impaired on metacognitive tasks relative to neurotypical controls. In contrast, both self-awareness-is-prior and two-systems theorists explicitly predict that metacognition should be unimpaired in ASD, and cite evidence from studies of ASD to support the idea of a dissociation between metacognition and mindreading in this disorder (see Goldman, 2006; Nichols & Stich, 2003). If such a dissociation between metacognition and mindreading could be established in the case of ASD, it would represent a major challenge to the one system account of the relation between (and origin of) these abilities. Therefore, one key aim of the current study was to test these predictions by investigating metacognition among autistic people using a classic, widely used metacognitive task (together with classic mindreading tasks).

A third source of evidence that has been used to evaluate the competing views in this debate comes from studies of metacognition in non-human primates (mostly macaques; see Beran, Smith, Coutinho, Couchman, & Boomer, 2009; Couchman, Beran, Coutinho, Boomer, & Smith, 2012; Smith, Beran, Couchman, & Coutinho, 2008; Smith, Beran, Redford, & Washburn, 2006; Smith, Couchman, & Beran, 2014; Smith, Redford, Beran, & Washburn, 2010; Smith, Shields, & Washburn, 2003; Washburn, Gulledge, Beran, & Smith, 2010). Based on the performance of macaques on measures designed to provide a behavioral index of metacognition, it has been claimed that monkeys are aware of their own states of uncertainty. That is, it has been claimed that macaques

[2] In the current paper we use both "autistic people" and "individuals with autism" to describe those diagnosed with ASD. We are aware that there are different opinions on whether identity first or person first is the best way when describing those with a diagnosis of ASD, but we aim to stay impartial in this debate. Therefore, we use both forms in order to remain balanced.

are capable of meta-representing their own knowledge/belief states and making strategic decisions on the basis of this meta-representation.[3]

The task that has probably been used most frequently to test metacognitive monitoring in non-human primates is the "strategic opt-out" uncertainty-monitoring task. In this paradigm, subjects are presented with a cognitive ("object-level") task that involves making a discrimination of some sort (e.g., between a densely versus sparsely pixelated stimulus, or requiring them to select the longest among lines of varying length). Successful discrimination results in a valued reward, while failure results in a penalty (loss of reward, including a "time out" before the next trial). The crucial feature of such paradigms, however, is that subjects have the opportunity to opt-out of any given trial, moving immediately to the next trial without penalty or reward. Participants who make adaptive use of the opt-out option, selecting it in cases where the discrimination is difficult (where they are otherwise likely to make a mistake), are said to be monitoring and responding to their own states of uncertainty. In other words, the individual is said to be meta-representing their own uncertainty about the correct response on difficult trials, with meta-representation driving adaptive behavioral responses. The basic finding in the comparative literature is that macaques, chimpanzees, and humans all make adaptive use of the opt-out option (showing very similar response profiles), while pigeons and capuchins generally do not (Smith et al., 2014).

Such claims are relevant to the debate about the relation between mindreading and metacognition, because there is no evidence that macaques are capable of meta-representing the epistemic states of conspecifics (Marticorena, Ruiz, Mukerji, Goddu, & Santos, 2011). If, therefore, this species is capable of meta-representing their own epistemic states, but not those of other macaques, then this would show the same dissociation between metacognition and mindreading that some have claimed is demonstrated by findings among autistic humans. If true, this would clearly represent a major challenge for the one system view and support the self-awareness-is-prior account or the two systems account. Such findings would also have methodological implications for the study of metacognition in humans, because they would provide a behavioral measure of the ability that could be employed with children or adults who have limited verbal communication ability and who could not therefore be tested using traditional metacognitive tasks.

The claim that such tasks really do reveal any sort of metacognitive awareness is by no means uncontroversial, however. Some critics have argued against a metacognitive interpretation of the uncertainty-monitoring paradigm by claiming that its findings can be explained in terms of mere associative learning, rather than metacognition (Le Pelley, 2012) (but see Smith et al., 2014, for a strong case against such associative learning models). Other critics of a metacognitive interpretation of the findings have claimed that the data can be explained in terms of first-order prospective, affect-involving, decision-making (Carruthers & Ritchie, 2012), of the sort that humans also engage in regularly (Gilbert & Wilson, 2007; Seligman, Railton, Baumeister, & Sripada, 2013). On this view, the animals in question *are* uncertain and this uncertainty drives behavioral responses. However, uncertainty is a first-order state of anxiety directed at the primary response options, resulting from an appraisal that the act of selecting either one of them is unlikely to succeed. As a result, those options *seem bad* (are negatively valenced), whereas the opt-out response is neutral or mildly positive. In order to select the latter, participants just need to pick the option that *seems best* in the circumstances; they don't need to be aware of, or meta-represent, their own state of uncertainty as such. In short, it has been claimed that while success in these tasks demonstrates cognitive decision-making, it fails to demonstrate metacognition.

In the current study, autistic and (closely-matched) non-autistic adults completed a strategic opt-out version of the uncertainty monitoring task that is frequently used in the comparative psychology literature, as well as a classic "judgement of confidence" task that is used widely among humans to measure metacognitive ability. In judgement of confidence tasks, participants make some form of perceptual or cognitive discrimination and then make a judgement about the likelihood that their discrimination was correct. The closer the correspondence between actual knowledge and judgements of knowledge, the better the individual's metacognitive accuracy. Investigating the performance of autistic individuals on the implicit opt-out alongside the traditional explicit judgement of confidence task provided an opportunity to test whether ASD involves a deficit in metacognition (indexed by diminished performance on the explicit judgement of confidence task) and whether the implicit opt-out task really measures metacognition. Finally, participants completed two classic tasks used to measure mindreading ability in humans. Use of mindreading tasks in the current study allowed us to investigate the extent to which implicit or explicit task performance relates to mindreading ability.

One aim of the study was to establish whether metacognition is really unimpaired in ASD, as some have claimed. There is a paucity of research into metacognitive monitoring in ASD and findings from studies of this ability have high relevance for clinical practice, as well as our understanding of the cognitive profile of strengths and difficulties in ASD more generally (as we discuss further in the General Discussion). A second aim was to explore the extent to which meta-representation of self (on the judgement of confidence task) was associated with meta-representation of others (on the mindreading tasks). A third aim was to test the claim that strategic opting-out of difficult trials on uncertainty-monitoring tasks *necessarily* requires some form of meta-representation. If implicit uncertainty-monitoring tasks tap some of the same meta-representational resources as explicit judgement-of-confidence tasks undoubtedly do, then performance on the two tasks should be related significantly. In contrast, if implicit uncertainty-"monitoring" is really a form of first-order affective decision-making, then performance in the two sorts of task should be uncorrelated, because only

[3] Notice that the claim here doesn't have to be that these tasks are tapping into explicit/conceptual forms of self-awareness, or explicit/conceptual forms of meta-representation. Indeed, comparative psychologists who interpret these tasks as tests of metacognition are apt to talk in terms of *degrees* or *gradations* of self-awareness among humans and animals (Smith et al., 2014). The implicit tasks in question might only require non-conceptual, perception-like, awareness of one's own states of uncertainty; but in humans the output of the same process can be conceptualized (and reported) *as* uncertainty. So there should still be significant overlap between the processes underlying these implicit uncertainty-monitoring tasks and those that underlie humans' explicit judgements of certainty. For explicit metacognitive judgments in humans will presumably be grounded in the sorts of nonconceptual awareness tested in implicit tasks, and only the latter might be available to non-human animals.

Table 1
Predictions made by the views under consideration.

Theory	Prediction				
	ASD *implicit* performance impaired	ASD *explicit* performance impaired	Implicit related to explicit	Mindreading related to *implicit*	Mindreading related to *explicit*
One system & implicit uncertainty monitoring is *not* metacognitive	No	Yes	No	No	Yes
Two systems & implicit uncertainty monitoring is *not* metacognitive	No	No	No	No	No
Self-awareness is prior & implicit uncertainty monitoring is *not*metacognitive	No	No	No	No	Yes[a]
One system & implicit uncertainty monitoring *is* metacognitive	Yes	Yes	Yes	Yes	Yes
Two systems & implicit uncertainty monitoring *is* metacognitive	No	No	Yes	No	No
Self-awareness is prior & implicit uncertainty monitoring *is* metacognitive	No	No	Yes	No	Yes[a]

[a] While mindreading ability should be predictive of explicit metacognitive performance on the self-awareness-is-prior account (as well as on the one-system view), it should be *less* predictive of metacognition among people with ASD. This is because many of the errors in mindreading among this population will result from damaged simulation abilities, which don't share resources with metacognitive ones. See the discussion in the text.

the judgement-of-confidence task is meta-representational.

In addition to their possible bearing on contrasting views of the relation between metacognition and mindreading, there is a second reason for investigating whether implicit uncertainty-monitoring tasks are genuinely metacognitive in nature. For if they are, this could provide researchers with an important and hitherto untapped tool for investigating the development of metacognitive abilities in young (even pre-verbal) human children. Moreover, such tasks could also provide a potentially important way of testing metacognitive abilities in nonverbal or minimally-verbal populations, including a significant proportion of people with ASD. We therefore set out to employ with our participants one of the implicit uncertainty-monitoring tasks commonly used in the comparative literature.

We thus seek to address two inter-related sets of questions. One concerns the relation between metacognition and mindreading; the other concerns the metacognitive status of implicit forms of uncertainty-monitoring. In the study described below, we set out to investigate both sets of questions in a single framework. We presented both implicit (non-verbal) and explicit (verbal) uncertainty-monitoring tasks to a group of adults with ASD and a group of age-, IQ-, and sex-matched neurotypical controls. We also measured mindreading abilities in both groups. We hoped that this design would enable us to discriminate between the three differing views of the relation between metacognition and mindreading (self-awareness-is-prior, two-systems, or one-system), while at the same time testing whether one of the implicit uncertainty-monitoring tasks that has commonly been employed with non-human animals is really a test of metacognitive ability.

The predictions of the various views outlined above are laid out in Table 1. Our own over-arching hypotheses were that (a) strategic opting-out on implicit (non-verbal) uncertainty-monitoring tasks does not require meta-representation, and (b) metacognition (as indexed by the judgement-of-confidence task) relies on the resources of the same meta-representational faculty as does mindreading. We therefore made the following more-detailed predictions for the outcome of the experiment.

(i) ASD-performance on the strategic opt-out task will *not* be impaired relative to the neurotypical performance. This is because we expect strategic opting-out to require only first-order representations, rather than meta-representations, and there is no reason (that we are aware of) to suggest that first-order valenced appraisals of stimuli would be challenging for intellectually able autistic adults.
(ii) ASD-performance on explicit tasks *will* be impaired relative to neurotypical performance. This is because these tasks require explicit metacognitive judgments, and because such judgments implicate the same resources as the mindreading faculty, which is widely believed to be compromised in ASD.
(iii) Implicit task performance (i.e., degree to which opting-out is adaptive/strategic) should not be predicted by metacognitive accuracy on the explicit judgement-of-confidence task in either group. This is because the two tasks do not tap into the same meta-representational resources. We expect the implicit task to require only first-order valenced decision making, whereas the explicit task requires meta-representation of one's own uncertainty.
(iv) Mindreading ability *will not* be related to implicit task performance. This is because the latter task is really a first-order one.
(v) Mindreading ability *will* be related to *explicit* metacognitive performance. This is because success in the explicit tasks requires self-awareness of one's own state of uncertainty, and because we expect self-awareness to implicate the same core meta-representational resources that are employed in mindreading.

These five predictions are laid out in the first row of Table 1. Subsequent rows represent the predictions of other possible combinations of views. Note that the first three rows represent the predictions of the three accounts of the relationship between

mindreading and metacognition on the assumption that implicit uncertainty-monitoring tasks are *not* metacognitive, whereas the bottom three rows represent those predictions on the assumption that uncertainty-monitoring *is* metacognitive. Note, too, that the predictions of the self-awareness-is-prior and two-system views are the same across both sets of assumptions, with the exception that only the former predicts that there should be a correlation between mindreading abilities and explicit metacognitive performance.

If the self-awareness-is-prior view predicts that mindreading and explicit metacognition should be correlated, and it is accepted that mindreading is damaged in ASD, then one might wonder why only the one-system account should predict that explicit metacognitive performance will be impaired in ASD (as we lay out in the second column of Table 1). But the self-awareness-is-prior account suggests that mindreading comprizes two distinct abilities: introspection/metacognition and imagination/simulation. The contribution of the former gives rise to the prediction that mindreading and explicit metacognition should correlate, but only the latter is thought to be damaged in ASD (Goldman, 2006). So there is no reason for a self-awareness-is-prior account to predict any impairment in metacognition in ASD. Proponents of the account could, of course, postulate independent damage to introspective abilities as well as to imaginative ones (impairment in which is thought to explain the absence of pretend play in childhood in ASD). But this is not a prediction made by self-awareness-is-prior accounts as currently formulated, and appears to lack any independent motivation.

2. Method

2.1. Participants

Twenty-one adults with ASD and 22 neurotypical comparison adults took part in the study. All participants completed the Wechsler Abbreviated Scale for Intelligence-II (Wechsler, 1999), which provides verbal, performance, and full-scale IQ scores. All participants also completed the Autism-spectrum Quotient (AQ; Baron-Cohen, Wheelwright, Skinner, Martin & Clubley, 2001. The AQ is a valid and reliable measure of ASD traits in people with a full diagnosis and in the general population. Participants read statements (e.g., “I find social situations easy”; “I find myself drawn more strongly to people than to things”) and decide the extent to which each statement applies to them, responding on a four-point Likert scale, ranging from “definitely agree” to “definitely disagree”. Scores range from zero to 50, with higher scores indicating more ASD traits and scores ≥ 26 representing clinically significant levels of ASD-like traits (Woodbury-Smith, Robinson, Wheelwright, & Baron-Cohen, 2005). In addition, participants with ASD completed the Autism Diagnostic Observation Schedule (ADOS, a detailed observational assessment of ASD features on which a score o ≥ 7 is consistent with a diagnosis of ASD (Lord et al., 2000). Participant characteristics and matching statistics are presented in Table 2, together with their scores on the mindreading tasks described in Section 2.2 below.

Participants in the ASD group had received verified diagnoses, according to conventional criteria (American Psychiatric Association, 2000; World Health Organization, 1993). No participant in either group reported current use of psychotropic medication or illegal recreational drugs, and none reported any history of neurological or psychiatric illness other than ASD.

2.2. Materials and methods

Mindreading task #1. The Reading the Mind in the Eyes Task (RMIE) (Baron-Cohen, Wheelwright, Hill, Raste, and Plumb, 2001) is a widely used measure of mindreading in clinical and non-clinical populations. Participants were presented with a series of 36 photographs of the eye region of the face. On each trial, participants were asked to pick one word from a selection of four to indicate what the person in the picture was thinking or feeling. If participants felt more than one of the words was applicable, they were instructed to select the word they thought was most suitable. Stimuli were presented on screen to participants in a random order, and no time limit was imposed. Scores on the RMIE task range from zero to 36, with higher scores indicating better performance on the task. The proportion of items correctly identified by ASD and comparison participants is shown in Table 2.[4]

Mindreading task #2. We also employed a version of the Animations Task (Abell, Happé, & Frith, 2000) as a second measure of mindreading. The task, which is based on Heider and Simmel (1944), required participants to describe interactions between a large red triangle and a small blue triangle, as portrayed in a series of silent video clips. Four such clips are apt to invoke an explanation of the triangles’ behavior in terms of epistemic mental states, such as belief, intention, and deception. These clips comprise the “mentalizing” condition of the task and were employed in this study.

Each clip was presented to participants on a computer screen. After the clip was finished, participants described what had happened in the clip. An audio recording of participants’ responses was made for later transcription. Each transcription was scored on a scale of zero to two for accuracy (including reference to specific mental states), based on the criteria outlined in Abell et al. (2000). Twenty percent of transcripts were also scored by two independent raters. Inter-rater reliability was excellent according to Cichetti

[4] It should be noted that although the RMIE can be characterized as a kind of empathy task, it is also undoubtedly a task that requires mindreading of the mental states of the target agents. For in each case what has to be selected is the most appropriate mental-state descriptor. The task has been employed in well over 250 studies, and shows good test-retest reliability (e.g., Fernández-Abascal, Cabello, Fernández-Berrocal, & Baron-Cohen, 2013), clearly distinguishes groups of participants with and without ASD (e.g., Baron-Cohen, Wheelwright, Hill, et al., 2001), is associated with the number of ASD traits shown by individuals in large population studies (e.g., Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), is correlated with other measures of mindreading even after the influence of IQ is controlled (e.g., Jones et al., 2018), and activates the same mindreading brain network as activated by other classic tasks that are assumed to measure mindreading (Schurz et al., 2014).

Table 2
Participant characteristics and mindreading performance.

	ASD (n = 21; 15 male)	NT (n = 22; 16 male)	t	p	d
Age (years)	37.15 (19.94)	37.21 (12.34)	−0.02	.99	< 0.01
VIQ	102.90 (11.72)	107.36 (9.62)	−1.37	.18	0.42
PIQ	99.67 (17.99)	105.41 (11.90)	−1.24	.22	0.38
FSIQ	101.05 (13.78)	106.86 (10.05)	−1.59	.12	0.48
AQ	32.14 (8.19)	15.77 (5.74)	7.62	< .001	2.02
ADOS	9.65 (4.80)	–			
RMIE (proportion)	0.67 (0.16)	0.76 (0.10)	−2.16	.036	0.65
Animations (proportion)	0.45 (0.21)	0.65 (0.23)	−3.04	.004	0.93
Mindreading composite (proportion)	0.56 (0.16)	0.71 (0.12)	−3.41	.001	1.03

VIQ = verbal IQ, PIQ = performance IQ, FSIQ = full scale IQ, AQ = Autism-spectrum Quotient (total score), ADOS = Autism Diagnostic Observation Schedule (social + communication score), RMIE = Reading the Mind in the Eyes test.

(1994) criteria (intra-class correlations > 0.82). Accuracy (proportion) among ASD and comparison participants is shown in Table 2.

Note that the Animations Task is a measure of spontaneous mental-state attribution. Those who score highly on this measure are spontaneously interpreting the movements of the geometric figures in the videos in mental-state terms (both affective and cognitive). Those who score low on this measure, in contrast, will tend to provide literal (non-mentalistic) descriptions of the movements they have observed.

In addition, to reduce the number of statistical comparisons made and thus reduce the chances of making a type I error, we created a composite mindreading score (shown in Table 2), which was created by averaging performance across the Animations and RMIE tasks, as the two tasks correlated significantly in the current sample (r = 0.32, p < .04).

The implicit strategic opt-out/uncertainty monitoring task. The implicit task was a perceptual discrimination task modeled closely on those that have been employed with non-human animals. On each trial, participants were presented with either two lines of differing length (in one version) or two patches of dots comprising differing numbers of dots (in the other version) (see Fig. 1a). In both versions, a red arrow was also presented to the right of the two stimuli (lines/patches). Participants were told that on each trial, they must click on one of the three response options within three seconds, and they were given a set of instructions outlining the consequences of each choice. They were instructed that they would start with a balance of £2.00, and would win or lose money depending on their responses. If the participant clicked on the longest line (or patch with greatest number of dots) they would gain 10p, whereas if they clicked on the shortest line (or patch with least number of dots) they would lose 10p. If they clicked on the red arrow, their score would remain the same and they would proceed immediately to the next trial. If they did not click on any of the options within 3 s, they would lose 30p.

Participants were asked whether they understood the rules, and were told that this was important because they would be asked to recall them at the end of the task. They were given a final chance to clarify the rules before beginning the task, but no further information about the nature of the task or the study aims was offered. Participants completed 10 practice trials on which their responses did not count toward their prize money, before completing 60 experimental trials on which their responses did count. At the end of the task, the participant's memory for the response rules was tested; the experimenter presented the participant with each rule and asked them to complete the payment amounts for each response type.

The order in which tasks (implicit/explicit) were completed was fixed, such that all participants completed the implicit task before they completed the explicit task. This was to minimize the chance that participants would adopt an explicit metacognitive strategy in solving the implicit tasks.

The primary dependent variable for object-level performance was the proportion of trials on which participants made accurate discrimination judgements. The primary dependent variable for "meta"-level performance was the difference between (a) the average perceptual discrimination difficulty on trials that participants opted out of making a discrimination judgment and (b) the average perceptual discrimination difficulty on trials that participants opted into making a discrimination judgment (i.e., trials on which they "took the test"). Trial difficulty was quantified as the similarity (in proportional terms) between the component items of each stimulus pair, with the smaller the value the more difficult the trial. For example, a trial on which the component items (length of the two lines, or number of dots in the two patches) differed by only 1% would be more difficult than a trial on which items differed by 10%. Evidence of adaptive opting-out would be indicated by participants opting out of trials on which the proportional difference between item pairs was significantly smaller than on trials that they opted into (e.g., opting-out of trials where there was only an average of 1% difference between stimulus pairs, but opting-into trials where there was an average of 10% difference between stimulus pairs).

For the purpose of correlation analyses, the average trial difficulty of trials that were opted out of was subtracted from the average trial difficulty of all trials that were opted-into. The larger and more positive the resulting value (an "adaptiveness score"), the more adaptive/strategic participants' opting-out was. If a participant never opted-out, the difference between options (opt-in/opt-out) could not be calculated. This was viewed as a lack of adaptive behavior, as participants did not appear to distinguish difficult from easy trials. Therefore, these participants were given a difference score of zero to reflect their lack of adaptive responding.

The explicit judgement of confidence meta*cognition task.* In the explicit task, participants completed the same perceptual discrimination task as they did during the implicit task, but with stimuli counterbalanced across tasks (those judging line length in the implicit task judged dot number in the explicit task, and vice versa) (see Fig. 1b). In the explicit task, however, the red arrow was not present, and participants were forced to choose between one of the two stimuli presented on the screen. On each trial, after a

Fig. 1a. provides a schematic illustration of the implicit task for both the lines and the dots conditions along with the visual feedback received. If participants clicked on the red arrow the trial was skipped and no feedback was received. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

discrimination judgement had been made, participants were instructed (on a new screen) to make a confidence judgment relating to their response on the perceptual discrimination task. Participants could select either "yes" or "no" to the question "Are you confident?" by clicking on the appropriate box within 3 s.

Participants started with a balance of £2.00 and were instructed that they would either win or lose money depending on their responses. If the participant clicked on the longest line (or patch with greatest number of dots) and clicked "yes" (indicating high confidence), they gained 10p, whereas if they clicked on the longest line and clicked "no" (indicating low confidence) they would gain nothing. If participants clicked on the shortest line (or patch with the fewest dots) and clicked "yes" (indicating high confidence) they would lose 10p, whereas if they clicked on the shortest line and clicked "no" (indicating low confidence) they would lose nothing. If they failed to respond on any trial within 3 s, they would lose 30p.

Participants were asked whether they understood the rules, and told that this was important because they would be asked to recall them at the end of the task. They were given a final chance to clarify the rules before beginning the task, but no further information about the nature of the task or about the study aims was offered. Participants completed 10 practice trials on which their responses did not contribute to their prize money, before completing 60 experimental trials on which their responses did contribute. At the end of the task, the participant's memory for the response rules was tested; the experimenter presented the participant with each rule and asked them to complete the payment amounts for each response type.

The dependent variable for object-level performance was the proportion of trials on which participants made the correct discrimination judgement. Gamma scores (Goodman & Kruskal, 1954) were calculated to provide an index of overall judgement of confidence meta-level accuracy. This analysis is recommended by Nelson (1984) and is commonly used to analyze judgement of confidence tasks (e.g., Kelemen, Frost, & Weaver, 2000; Nelson & Narens, 1990; Nelson, Narens, & Dunlosky, 2004; Wojcik, Moulin, & Souchay, 2013). Gamma scores are a measure of association (between meta-level judgements of performance and actual object-level performance) and were calculated by comparing the number of correct meta-judgements with the number of incorrect judgements made by each individual. To calculate gamma scores the formula $G = (ad - bc)/(ad + bc)$ was used, with "a" representing the number of "Yes" (confident) judgements made following a correct perceptual discrimination, "b" the number of incorrect "Yes" (confident) judgements made following an incorrect perceptual discrimination, "c" the number of incorrect "No" (not confident) judgements following a correct perceptual discrimination, and "d" the number of correct "No" (not confident) judgements following an incorrect perceptual discrimination. Gamma scores range between $+1$ to -1, where a score of 0 indicates chance-level accuracy, positive values indicate above chance accuracy, and negative values indicate below chance accuracy. When calculating gamma scores, the score cannot be calculated when two or more of the prediction rates (a, b, c, or d) are equal to 0. As such, the raw data were adjusted by adding 0.5 onto each prediction frequency and dividing by the overall number of judgement of confidence judgements made (N) plus 1 ($N + 1$). This correction is recommended by Snodgrass and Corwin (1988) and is routinely used when calculating gamma scores on metamemory tasks (e.g., Bastin et al., 2012; Wojcik et al., 2013).

2.3. Statistical analyses

A standard alpha level of 0.05 was used to determine statistical significance. All reported significance values are for two-tailed tests, unless reported otherwise for specific, predicted results. Where ANOVAs were used, we report η_p^2 values as measures of effect

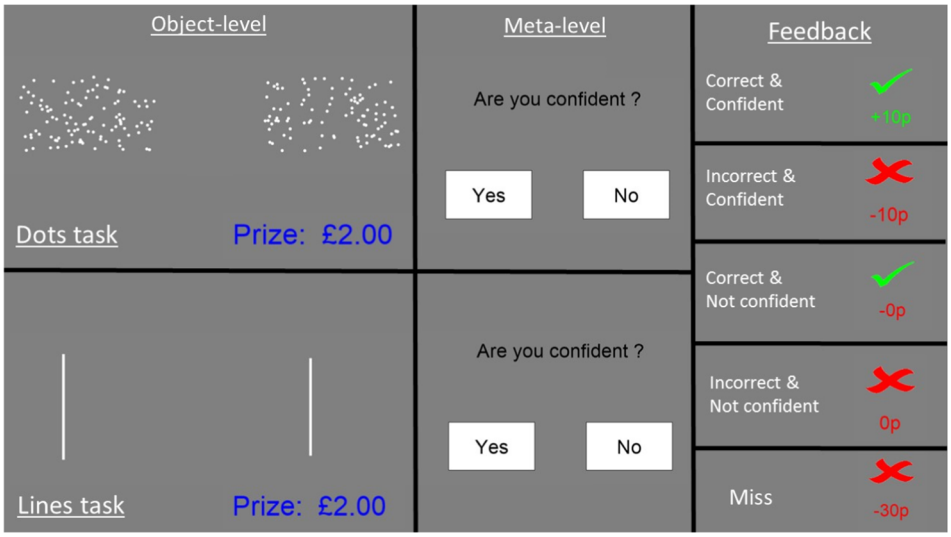

Fig. 1b. provides a schematic illustration of the explicit task for both the lines and the dots conditions along with the meta-level question they received immediately afterwards and the visual feedback they received depending upon their object level accuracy and the meta-level confidence.

size (≥0.01 = small effect, ≥0.06 = moderate effect, ≥0.14 = large effect; Cohen, 1969). Where t-tests were used, we report Cohen's *d* values as measures of effect size (≥0.20 = small effect, ≥0.50 = moderate effect; ≥0.80 = large effect; Cohen, 1969).

3. Results

3.1. Implicit task performance

With regard to object-level performance, there were no significant differences between participants with ASD ($M = 0.70$, $SD = 0.08$) and comparison participants ($M = 0.72$, $SD = 0.07$) in the proportion of trials where stimuli were correctly discriminated, $t(41) = 1.20$, $p = .24$, $d = 0.04$. Thus, the two groups were very similar with respect to cognitive-level ability. Moreover, there was no significant between-group difference in the number of payment rules recalled, $t(41) = 0.63$, $p = .53$, $d = 0.19$. Finally, the difference between ASD participants ($M = 0.94$, $SD = 0.10$) and comparison participants ($M = 0.87$, $SD = 0.17$) in the proportion of trials that were opted into was non-significant (albeit moderate in size), $t(41) = 1.69$, $p = .10$, $d = 0.50$.

With regard to meta-level performance, the average difference in proportional similarity of stimulus pairs on trials opted into was compared to the average difference in proportional similarity of stimulus pairs on trials opted out of, among each diagnostic group. Data were subjected to a 2 (Group: ASD/neurotypical) × 2 (Condition: opt-in/opt-out) mixed ANOVA. The ANOVA yielded a significant main effect of Condition, reflecting greater discrimination difficulty on trials that were opted-out of than trials that were opted-into, $F(1, 41) = 21.32$, $p < .001$, η_p^2=0.34. However, neither the main effect of Group, $F(1, 41) = 1.18$, $p = .28$, η_p^2=0.03, nor the Group × Condition interaction, $F(1, 41) = 0.92$, $p = .34$, η_p^2=0.02, was significant, and both were associated with small effect sizes. In other words, there were no significant differences between the diagnostic groups in terms of either levels or patterns of performance across conditions, with both showing similarly adaptive opting into and out of trials. Fig. 2 displays these results. Trials that autistic participants opted out of were significantly more difficult than the trials they opted into, reflecting adaptive opting out within this group, $t(20) = 2.90$, $p = .009$, $d = 0.67$. Likewise, the trials that neurotypical participants opted out of were significantly more difficult than the trials they opted into, $t(21) = 3.63$, $p = .002$, $d = 0.76$. Most important, there was no significant between-group difference in the degree to which opting out was adaptive/strategic, $t(41) = 0.99$, $p = .33$, $d = 0.31$, as reflected by the non-significant interaction effect in the ANOVA and as shown in Fig. 2. To ensure that the "meta-" level results were not being influenced by general cognitive-level performance, object-level discrimination accuracy, number of payment rules correctly recalled, and proportion of trials that were opted out of were covaried in a further analysis of covariance.[5] None of these covariates had a significant influence on degree of adaptive opting out (all *p*s > .12, all η_p^2<0.06) and, most important, the effect of Group remained non-significant, $F(1, 38) = 0.19$, $p = .67$, η_p^2=0.005.

[5] It is important in case-control studies of metacognition that the participant groups are equated for cognitive-level task variables, because cognitive-level task performance is involved in the calculation of metacognitive accuracy. Thus, ensuring that (a) groups are equated for cognitive-level performance, and (b) cognitive-level performance is controlled ensures that analyses of between-group differences in "meta"-level performance are not contaminated by group differences in cognitive-level performance.

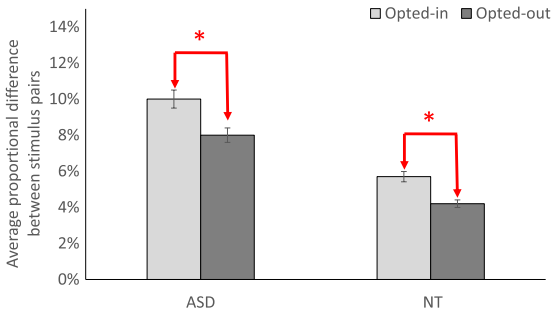

Fig. 2. illustrates the average proportional difference between stimuli for both opted-out and opted-in trials among participants from each diagnostic group for the implicit task. *p < .01 for within-group contrasts.

3.2. Explicit task performance

With regard to object-level performance, there was no significant difference between participants with ASD ($M = 0.69$, $SD = 0.07$) and comparison participants ($M = 0.70$, $SD = 0.07$) in the proportion of trials on which stimuli were correctly discriminated, $t(41) = 0.38$, $p = .71$, $d = 0.14$. Thus, the two groups were very similar with respect to cognitive-level ability. Moreover, there was no significant between-group difference in the number of payment rules recalled, $t(41) = 1.44$, $p = .16$, $d = 0.43$. Finally, the difference between ASD participants (M = 0.80, SD = 0.16) and comparison participants (M = 0.76, SD = 0.13) in the proportion of confident judgements made was non-significant, $t(41) = 0.93$, $p = .36$, $d = 0.27$.

With regard to meta-level performance, Fig. 3 shows the mean gamma score among participants from each diagnostic group. Gamma was significantly smaller among autistic participants than NT participants, $t(41) = 2.21$, $p < .03$, $d = 0.67$, reflecting a moderate-to-large diminution of metacognitive monitoring accuracy in ASD. To ensure that between-group differences in metacognitive accuracy were not being driven by differences in general cognitive-level performance, object-level discrimination accuracy, number of payment rules correctly recalled, and proportion of confident responses made were covaried in a further analysis of covariance. None of these covariates had a significant influence on gamma score (all $ps > .21$, all $\eta_p^2 = < 0.04$), but the effect of Group was significant, $F(1, 38) = 5.47$, $p < .03$, η_p^2=0.13. Thus, autistic participants had a significant diminution of metacognitive accuracy (associated with a borderline-large effect size) even after the influence of all other task variables had been controlled (note: an η_p^2 of 0.13 is equivalent to a Cohen's d of approximately 0.78).

3.3. Relation between implicit and explicit "meta"-level task performance

An initial correlation analysis indicated that the degree of strategic opting-out on the implicit uncertainty monitoring task was non-significantly associated with gamma score on the judgement-of-confidence task, $r = 0.12$, $p = .43$. In order to explore the extent to which performance on the explicit task predicted performance on the implicit task, a further regression analysis was conducted. The implicit task adaptiveness score was the *dependent variable* and judgement-of-confidence gamma score was used as the *predictor variable*. A Group × Gamma interaction variable was also included to establish whether judgement of confidence was a predictor of implicit adaptiveness score among one group of participants only. If strategic opting-out depends on meta-representation of own uncertainty, then meta-level performance on the judgement-of-confidence task (which clearly requires meta-representation of one's

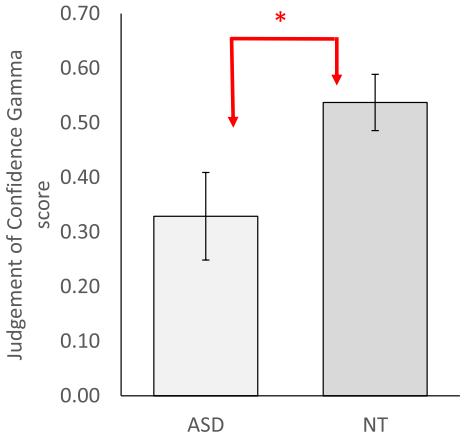

Fig. 3. illustrates the mean gamma score among participants from each diagnostic group for the explicit task. *p < .03 for between group comparison.

Table 3
Summary of regression analyses of the relation between adaptive performance on the implicit task and adaptive performance on the explicit task.

	B	*SE B*	β	*t*	*p*
DV: Implicit adaptiveness difference score[a]					
Block 1					
Explicit Gamma score	0.007	0.009	0.12	0.80	.43
Block 2					
Explicit Gamma score	0.01	0.01	0.21	1.17	.25
Explicit Gamma score × Group	0.01	0.01	0.17	0.97	.34

[a] $R^2 = 0.02$ for Block 1 ($p = .43$); $\Delta R^2 = 0.02$ for Block 2 ($p = .34$).

own mental states) should predict the degree to which opting out was performed strategically. The results are shown at the top of Table 3.

In Block 1, judgement of confidence Gamma was a non-significant predictor of implicit adaptiveness score, accounting for only 2% of variance in the latter. Gamma remained a non-significant predictor in Block 2 and the new Group × Gamma interaction-variable was also non-significant. Thus, the extent to which participants opted out strategically on the implicit task was not predicted significantly by their metacognitive accuracy on the explicit task among either group of participants.

3.4. Relations between mindreading, judgement-of-confidence gamma, and adaptiveness of opting out

In order to explore the relation between adaptive performance on the implicit and explicit tasks, on the one hand, and mind-reading ability, on the other, two further regression analyses were conducted (Table 4). In the first analysis, the *implicit* adaptiveness difference-score was the dependent variable and the composite mindreading score was the predictor variable. In addition, a Group × mindreading interaction-variable was included to establish whether the extent to which mindreading was a predictor of the dependent variable differed across groups. In Block 1, the mindreading composite score was a non-significant predictor of implicit adaptiveness difference-score, accounting for < 1% of variance in the latter. The mindreading composite remained a non-significant predictor in Block 2 and the new Group × mindreading interaction variable was also non-significant. Thus, mindreading ability was a non-significant predictor of the extent to which participants made adaptive responses on the implicit task and there was no significant between-group difference in this extent.

In the second analysis, Gamma on the explicit task was the dependent variable and the composite mindreading score was the predictor variable. Again, a Group × mindreading interaction variable was included. In Block 1, the mindreading composite score was a significant predictor of explicit Gamma, accounting for 14% of variance in the latter. The mindreading composite score remained a significant predictor in Block 2, but the new Group × mindreading interaction variable was non-significant. Thus, mindreading ability was a significant predictor of the extent to which participants made accurate metacognitive judgements of confidence and there was no significant between-group difference in this extent.[6]

4. Discussion

We employed an implicit opt-out task that was modeled closely on measures of (alleged) metacognitive ability widely employed with animals, as well as an explicit judgement-of-confidence task that is widely used with humans to measure metacognitive accuracy. Findings were closely in accord with our predictions. First, they indicate that autistic individuals have impaired metacognitive monitoring ability, as reflected by their diminished meta-level accuracy on the judgement-of-confidence task (a classic task, used widely to measure metacognitive ability among humans). Thus, autistic participants had significant difficulties both with meta-representation of others (they were impaired on each of the measures of mindreading undertaken) and with meta-representation of self (diminished judgement-of-confidence accuracy). This is contrary to the suggestion made by some that metacognition is intact in ASD and that that the disorder shows a dissociation between (intact) metacognition and (impaired) mindreading. The between-group difference in judgement-of-confidence accuracy was borderline large after controlling for other task variables ($d = 0.78$), and is comparable to the size of mindreading impairment typically observed among autistic individuals ($d = 0.88$; Yirmiya et al., 2001) and observed among autistic participants in the current sample ($d = 0.65$ for RMIE and 0.93 for Animations). This is compelling evidence that meta-representation of self is impaired in ASD and that this impairment is of a magnitude that is comparable to the known impairment of mindreading in this disorder.

The finding of impaired metacognition in ASD is predicted by the view that metacognition and mindreading share a set of meta-representational resources. However, it is problematic for all alternative views of the relation between mindreading and metacognition, unless one were to postulate that a dedicated metacognitive system is independently damaged in ASD, in addition to the

[6] The judgement of confidence Gamma score was associated significantly with both Animations performance, $r = 0.35$, $p = .02$, and RMIE performance, $r = 0.26$, $p = .04$ (one-tailed). We also ran separate regression analyses with the RMIE and Animation tasks individually. While RMIE alone did not predict judgement of confidence accuracy after the Group x gamma interaction was included, the Animations task did significantly predict judgement of confidence gamma in all blocks.

Table 4
Summary of regression analyses of the relation between mindreading, and adaptive performance on the implicit and explicit tasks.

	B	*SE B*	β	*t*	*p*
DV: Implicit adaptiveness difference-score[a]					
Block 1					
Mindreading composite score	0.01	0.02	0.09	0.56	.58
Block 2					
Mindreading composite score	0.01	0.02	0.09	0.56	.58
Mindreading composite score × group	0.01	0.009	0.23	1.47	.15
DV: Explicit Gamma-score[b]					
Block 1					
Mindreading composite score	0.78	0.30	0.38	2.66	**.01**
Block 2					
Mindreading composite score	0.73	0.30	0.36	2.45	**.02**
Mindreading composite score × group	0.16	0.16	0.15	1.00	.33

[a] $R^2 = 0.008$ for Block 1 (p = .58); $\Delta R^2 = 0.05$ for Block 2 ($p = .15$).
[b] $R^2 = 0.15$ for Block 1 (p = .01); $\Delta R^2 = 0.17$ for Block 2 ($p = .33$).

mindreading system that is widely-acknowledged to be impaired among people with this disorder. Yet none of the theorists in this debate have *predicted* such a "double hit". Indeed, theorists who reject a one-system view are generally clear that they predict people with ASD should have *preserved* metacognitive abilities (Goldman, 2006; Nichols & Stich, 2003). Given the importance in driving scientific progress of disproving predictions, this finding of impaired metacognition is important for theory building.

Moreover, our finding that mindreading abilities were predictive of explicit metacognitive performance on the judgement-of-confidence task suggests that mindreading and metacognition are not underpinned by two distinct mechanisms, as Nichols and Stich (2003) claim. Taken on its own, however, this result is consistent with Goldman (2006) self-awareness-is-prior account, since on his view mindreading abilities depend on metacognitive ones (together with simulation ability). But in fact our data raise a problem for the latter account also. For if, as Goldman assumes, it is simulation abilities that are damaged in individuals with ASD, then it will be these weaker simulation abilities that are responsible for the poorer mindreading performance of this group. That means that a significant proportion of the mindreading deficit in autistic individuals will result from an ability that is plays no part in metacognitive performance; namely, simulation ability. And in that case we should expect mindreading abilities to be *less* predictive of metacognitive abilities in autistic than in neurotypical individuals. But we found no evidence that this is so.

These findings (that metacognition is impaired in ASD and that metacognitive ability is predicted by mindreading ability) provide support for one-system views of the relation between metacognition and mindreading. A challenge for one-system views, however, has been to explain the findings that have been taken to indicate that some non-human primates are capable of metacognition, but not mindreading. Our results are relevant to this issue too. In the present study, autistic participants had clear impairments in meta-representing self (on the judgement-of-confidence task) and others (on the RMIE and Animations tasks), but yet still opted out of implicit-confidence trials in the same strategic way that age- and IQ-matched neurotypical comparison participants did. The between-group difference in the degree to which opting out was adaptive/strategic was statistically negligible/small and non-significant, which contrasts with the significant and borderline large impairments in both mindreading (on the RMIE and Animations tasks) and metacognitive accuracy (on the judgement-of-confidence task) observed among autistic participants. Moreover, adaptiveness of opting out was not predicted by either metacognitive ability or mindreading ability; all associations with opting out were negligible in size.

These findings suggest that opting out of difficult trials doesn't *necessarily* require meta-representation (nor does it actually involve meta-representation in humans who undertake these tasks), but can be achieved using risk-based affective appraisals of likely success. Of course, it does not *prove* beyond doubt that non-human primates do not meta-represent their own states when they opt-out of difficult trials on equivalent versions of uncertainty monitoring tasks.[7] Rather, it shows that it need not be the case that they strategically opt out by meta-representing self. In the current study, we observed that autistic participants had statistically significant and large difficulties with meta-representing self and others, yet nonetheless opted out of difficult trials strategically. Combined with the findings that the degree to which participants opted out strategically was not predicted by either metacognitive ability or mindreading ability (whereas mindreading ability did predict explicit metacognitive ability), we suggest that a high degree of caution should be taken when interpreting any findings of strategic opting out as indicating meta-representation of self. These findings are relevant to theory-building, of course, but also have methodological implications in that they suggest implicit uncertainty-monitoring

[7] There are, of course, many different forms of metacognition, and many different non-verbal paradigms have been employed with non-human animals by comparative researchers. These include information-seeking paradigms and memory-monitoring paradigms, in addition to uncertainty-monitoring. Here we focus just on the latter. We make no claim about the metacognitive credentials of the other paradigms. Nor, of course, do we argue here that macaques are actually incapable of metacognition; nor even do we claim that that they aren't using metacognition when they respond adaptively in uncertainty-monitoring tasks. Rather, we present evidence that calls into question the idea that the implicit uncertainty-monitoring paradigm *necessarily* provides a test of metacognitive ability.

tasks cannot be used to reveal nascent self-awareness abilities in non-verbal or minimally-verbal human populations.

Couchman et al. (2012) point out, however, that when humans are queried after undertaking an implicit uncertainty-monitoring task of the sort employed here, they report that they opted out when they did because they were uncertain of the correct response. Isn't this evidence that people respond as they do in these tasks because they are aware of their own uncertainty? But it is one thing to say that humans (and macaques) opt out because they are uncertain (which we agree with), and quite another thing to say that they opt out *because* they are *aware* of being uncertain. Since humans are consummate mindreaders, they will of course correctly judge that they responded as they did because they were uncertain. As such, they retrospectively explain their own behavior by appealing to epistemic mental states, which requires mindreading. However, this doesn't show that they were concurrently aware of their uncertainty *while* they were responding, nor that they responded as they did *because of* such awareness. We submit that if humans sometimes make the latter sort of causal claim, they go beyond any evidence that *could* have been available to them at the time. (Indeed, no one should think that causal relations among mental states are introspectable.) Moreover, humans are known to engage in prospective valence-based decision-making in conditions of uncertainty (Gilbert & Wilson, 2007; Seligman et al., 2013). Indeed, valence is increasingly seen as the "common currency" for deciding among varying values under conditions of risk (Levy & Glimcher, 2012; Ruff & Fehr, 2014). Note that these models of human decision-making are *not* metacognitive ones.

If it is granted that implicit uncertainty-monitoring isn't genuinely metacognitive, then when considering the respective predictions of one-system ("metacognition and mindreading are grounded in a common meta-representational faculty") and multiple-system views of the relationship between mindreading and metacognition, the only relevant rows in Table 1 are the first three. So the relevant differences are just that (1) the one-system view predicts that explicit metacognitive performance should be weaker in participants with ASD (who are known to be weaker at mindreading), whereas both the two-system account and the self-awareness-is-prior account do not; (2) the one-system view predicts that mindreading ability should be predictive of explicit metacognitive performance, whereas the two-system account does not; and (3) the one-system view predicts there should be no differences between groups in the extent to which mindreading ability predicts metacognitive performance, whereas the self-awareness-is-prior account predicts an interaction. Our findings support the one-system view in each case.

If the one-system view is correct, however, then why is it that mindreading ability explains only 14% of the variance in explicit metacognitive performance? If metacognition and mindreading share a common meta-representational faculty, then one might expect that variability in mindreading ability should account for a larger proportion of the variance in metacognition than this. In reply, it is important to note it would be implausible to claim that mindreading is a monolithic system. Rather, it is likely to comprise a core set of resources (or "core knowledge"; Spelke & Kinzler, 2006), combined with attribution procedures and inference rules specific to certain kinds of mental state and/or certain sorts of task or situation. It seems likely, for example, that the mindreading resources required to identify an emotional expression from someone's eyes (as required to succeed on the Reading the Mind in the Eyes Task) only partially overlap with those accessed when spontaneously interpreting the movements of seemingly-animated geometrical shapes on a screen (as required for the Animations Task). And then even on a one-system account, according to which metacognition and mindreading rely on the same faculty, many of the principles and resources involved in explicit forms of metacognition are likely to be different again. So it is not surprising that the measures of mindreading we employed should only correlate with explicit metacognitive performance to a limited degree, especially since the tasks differ from one another in many other respects (for example, in their executive demands).

It is also important to note that this finding is a replication of a recent one by Williams, Bergstrom, and Grainger (2016), who found that the accuracy of explicit metacognitive judgements of confidence was associated significantly with mindreading ability in a large sample of neurotypical individuals. Thus, although mindreading ability is predictive of metacognitive judgement accuracy to a relatively modest degree, the association would appear to be a reliable one.

Taken together, our findings are in keeping with only one of the theories considered, which is striking given that several predictions overlap across theories, highlighting how challenging it is to distinguish them empirically. The findings suggest, on the one hand, that traditional implicit uncertainty-monitoring paradigms of the sort used to test "metacognition" in non-human animals do not, in fact, measure awareness of one's own mental states (or at any rate, not in humans). On the other hand, the results suggest that accuracy of explicit metacognitive judgements about one's own mental states depends to a significant extent on mindreading ability. As a result, people with ASD, who have established mindreading difficulties, also show significantly weaker performance on explicit metacognitive tasks (but not implicit uncertainty-monitoring ones).

The fact that we predicted exactly this pattern of results may be surprising to some autism researchers in the field, given that, in domains other than metacognition, explicit task performance is sometimes less impaired among autistic individuals than performance on implicit tasks (see, for example, Brosnan & Ashwin, 2018). There are various explanations for this pattern of performance (i.e., explicit superior to implicit) when observed among participants with ASD, such as compensatory strategy use on explicit tasks or alternative reasoning styles. However, these explanations rely on an underlying assumption that the implicit task in question depends on the same conceptual resources as the explicit task. When planning the current study we assumed, quite the contrary, that the implicit opt-out task did not rely on the same metacognitive resources as the explicit judgement of confidence task. Thus, the prediction that opt-out task performance would be undiminished in ASD was not out of keeping with findings in other domains. Of course, the question of whether intellectually high-functioning autistic adults would employ compensatory strategies to perform well on the explicit task, despite limited underlying metacognitive competence, was an open one. In the current study, this was clearly not the case, but we would not rule out the possibility that other studies could observe such a finding. The main conclusion from the current study is that there was a clear differentiation between the implicit and explicit tasks both in terms of levels of performance observed among autistic participants and in terms of associations with other cognitive tasks. This is in keeping only with the set of predictions that we made prior to beginning the study.

In addition, our findings have implications for theoretical topics beyond those that have been our main focus here. They suggest, for example, that one shouldn't uncritically accept metacognitive interpretations of adaptive responding on implicit uncertainty-monitoring tasks, not only when participants are non-human animals, but also when they are pre-verbal human infants (e.g., Goupil, Romand-Monnier, & Kouider, 2016). This is not to suggest that preverbal infants are incapable of meta-representing their own states of confidence or uncertainty. (After all, they do appear to be capable of representing others' states of mind; e.g, Baillargeon, Scott, & He, 2010.) Rather, we are suggesting that care should be taken when selecting an experimental technique to demonstrate self-awareness in infants. One needs to be satisfied that the task taps into more than just first-order cognitive processes; and on our view, implicit uncertainty-monitoring tasks do not fit the bill.

From a clinical perspective, these results suggest that intervention efforts designed to remediate mindreading deficits in ASD should have an added positive influence on metacognitive monitoring abilities among people with this disorder. Equally, however, the finding that the overlap between metacognition and mindreading was not absolute in the current study suggests that individuals with ASD would also benefit from targeted training in metacognitive monitoring. Previous research has shown that metacognitive monitoring predicts educational outcomes independent of general cognitive ability (Hartwig & Dunlosky, 2012; Thiede, 1999; Veenman, Wilhelm, & Beishuizen, 2004), and cognitive deficits can be improved by employing metacognitive strategies (Dunlosky, Kubat-Silman, & Hertzog, 2003; Murphy, Schmitt, Caruso, & Sanders, 1987). This is of particular relevance for individuals with ASD, many of whom experience significantly lower academic achievement based on their level of intelligence than would be expected, which impacts their life chances negatively (Estes, Rivera, Bryan, Cali, & Dawson, 2011). Therefore, training in metacognitive monitoring has the potential to improve both educational attainment and quality of life among individuals with ASD.

Acknowledgements

The authors would like to thank all of the participants who took part in this study. The authors would also like to the Kent Autistic Trust for assistance with participant recruitment. Without the support of these people and institutions, this research would not have been possible. This research was funded by an Economic and Social Research Council Research Grant (ES/M009890/1) awarded to David Williams, Sophie Lind, and Peter Carruthers.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.concog.2019.01.013.

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.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revised) (DSM-IV-TR)Washington DC: American Psychiatric Association.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington DC: American Psychiatric Association.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Science, 14*(3), 110–118.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001a). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241–252.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001b). 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.

Bastin, C., Feyers, D., Souchay, C., Guillaume, B., Pepin, J. L., Lemaire, C., ... Salmon, E. (2012). Frontal and posterior cingulate metabolic impairment in the behavioral variant of frontotemporal dementia with impaired autonoetic consciousness. *Human Brain Mapping, 33*(6), 1268–1278.

Beran, M., Smith, J., Coutinho, M., Couchman, J., & Boomer, J. (2009). The psychological organization of "uncertainty" responses and "middle" responses: A dissociation in capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes, 35*, 371–381.

Brosnan, M., & Ashwin, C. (2018). Reasoning on the autism spectrum. *Encyclopedia of autism spectrum disorders* (pp. 1–7). Springer.

Brunsdon, V. E. A., & Happé, F. (2014). Exploring the 'fractionation' of autism at the cognitive level. *Autism, 18*(1), 17–30.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32*, 121–138.

Carruthers, P. (2011). *The opacity of mind.* Oxford University Press.

Carruthers, P., & Ritchie, J. B. (2012). The emergence of metacognition: Affect and uncertainty in animals. In M. Beran, J. Brandl, J. Perner, & J. Proust (Eds.). *Foundations of metacogntion.* Oxford University Press.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125*(8), 1839–1849.

Cichetti, D. V. (1994). Guidelines, criteria, and rule of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences.* New York: Academic Press.

Couchman, J., Beran, M., Coutinho, M., Boomer, J., & Smith, J. D. (2012). Evidence for animal metaminds. In M. Beran, J. Brandl, J. Perner, & J. Proust (Eds.). *Foundations of metacogntion.* Oxford.

Descartes, R. (1637). The Discourse on the Method. Many editions and translations are available.

Descartes, R. (1641). Meditations on First Philosophy. Many editions and translations are available.

Dunlosky, J., Kubat-Silman, A. K., & Hertzog, C. (2003). Training monitoring skills improves older adults' self-paced associative learning. *Psychology and Aging, 18*(2), 340.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Sage Publications.

Estes, A., Rivera, V., Bryan, M., Cali, P., & Dawson, G. (2011). Discrepancies between academic achievement and intellectual ability in higher-functioning school-aged children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 41*(8), 1044–1052.

Fernández-Abascal, E. G., Cabello, R., Fernández-Berrocal, P., & Baron-Cohen, S. (2013). Test-retest reliability of the 'Reading the Mind in the Eyes' test: A one-year follow-up study. *Molecular Autism, 4*(1), 33.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906.

Gilbert, D., & Wilson, T. (2007). Prospection: Experiencing the future. *Science, 317*, 1351–1354.
Goldman, A. (2006). *Simulating minds.* Oxford University Press.
Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American statistical association, 49*(268), 732–764.
Gopnik, A. (1993). The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16*, 1–14.
Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences, USA, 113*, 3492–3496.
Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*(1), 126–134.
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243–259.
Jones, C. R., Simonoff, E., Baird, G., Pickles, A., Marsden, A. J., Tregay, J., ... Charman, T. (2018). The association between theory of mind, executive function, and the symptoms of autism spectrum disorder. *Autism Research, 11*(1), 95–109.
Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28*(1), 92–107.
Le Pelley, M. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 686–708.
Levy, D., & Glimcher, P. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology, 22*, 1027–1038.
Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... Rutter, M. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223.
Marticorena, D., Ruiz, A., Mukerji, C., Goddu, A., & Santos, L. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science, 14*, 1406–1416.
Murphy, M. D., Schmitt, F. A., Caruso, M. J., & Sanders, R. E. (1987). Metamemory in older adults: The role of monitoring in serial recall. *Psychology and Aging, 2*(4), 331.
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109.
Nelson, T., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Vol. Ed.), *The psychology of learning and information: Vol. 26*Academic Press.
Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment Recall and Monitoring (PRAM). *Psychological Methods, 9*(1), 53.
Nichols, S., & Stich, S. (2003). *Mindreading.* Oxford University Press.
Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? *Behavioral and Brain Sciences, 1*(4), 592–593.
Ruff, C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision-making. *Nature Reviews Neuroscience, 15*, 549–562.
Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9–34.
Seligman, M., Railton, P., Baumeister, R., & Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science, 8*, 119–141.
Smith, J. D., Beran, M., Couchman, J., & Coutinho, M. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review, 15*, 679–691.
Smith, J. D., Beran, M., Redford, J., & Washburn, D. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General, 135*, 282–297.
Smith, J. D., Couchman, J., & Beran, M. (2014). Animal metacognition: A tale of two comparative psychologies. *Journal of Comparative Psychology, 128*, 115–131.
Smith, J. D., Redford, J., Beran, M., & Washburn, D. (2010). Rhesus monkeys (*Macaca mulatta*) adaptively monitor uncertainty while multi-tasking. *Animal Cognition, 13*, 93–101.
Smith, J. D., Shields, W., & Washburn, D. (2003). The comparative psychology of uncertainty monitoring and meta-cognition. *Behavioral and Brain Sciences, 26*, 317–373.
Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34.
Spelke, E., & Kinzler, K. (2006). Core knowledge. *Developmental Science, 10*, 89–96.
Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*(4), 662–667.
Veenman, M. V., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14*(1), 89–109.
Washburn, D., Gulledge, J., Beran, M., & Smith, J. D. (2010). With his memory magnetically erased, a monkey knows he is uncertain. *Biology Letters, 6*, 160–162.
Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence.* New York, NY: The Psychological Corporation: Harcourt Brace & Company.
Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*, 655–684.
Williams, D. M., Bergstrom, Z., & Grainger, C. (2016). The hypercorrection effect in children with ASD: Evidence of impaired metacognition? *Autism: International Journal of Research and Practice*. https://doi.org/10.1177/1362361316680178.
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.
Woodbury-Smith, M. R., Robinson, J., Wheelwright, S., & Baron-Cohen, S. (2005). Screening adults for asperger syndrome using the AQ: A preliminary study of its diagnostic validity in clinical practice. *Journal of Autism and Developmental Disorders, 35*, 331–335.
World Health Organization (1993). *International classification of mental and behavioral disorders: Clinical descriptions and diagnostic guidelines* (10th ed.). Geneva, Switzerland: World Health Organization.
Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin, 124*(3), 283–307. https://doi.org/10.1037//0033-2909.124.3.283.