Cognition 225 (2022) 105117

Contents lists available at ScienceDirect

Cognition

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

Model-free metacognition

Peter Carruthers [a,*], David M. Williams [b]

[a] Department of Philosophy, University of Maryland, College Park, MD 20742, USA
[b] School of Psychology, University of Kent, Keynes College, Canterbury, Kent CT2 7NP, UK

ARTICLE INFO

Keywords:
Curiosity
Ignorance
Mental effort
Meta-representation
Uncertainty

ABSTRACT

Extensive work has been done on the metacognitive capacities of humans, as well as to investigate metacognitive processes in nonhuman animals. What we propose here, however, is that there are two very different forms that metacognition can take: either model-based (implicating at least a simplified model of the thinker's own mind), or model-free (representing some mental state or process in oneself in the absence of any such model). The focus of all work on human metacognitive judgments has been on the model-based variety, as have been most attempts to discover metacognition in animals. We first review recent studies suggesting that there are no resources shared between human metacognitive judgments and the sorts of behavioral tests employed with animals, implying that the latter fail to provide evidence of even simplified forms of model-based metacognition. Thereafter the question of model-*free* metacognition in animals is pursued. Negative verdicts are rendered on a pair of possible claims of this sort. But two positive answers are defended. One is that epistemic emotions like curiosity and interest, as well as the signals involved in failed memory searches, implicate representations whose content is, *unknown*. The other is that decisions to deploy attentional / mental effort (which many animals besides humans can do) depend on appraisals of an analog-magnitude signal representing the extent to which executive resources are engaged.

1. Introduction

Metacognition is defined in the field as "thinking about thinking" (Dunlosky & Metcalfe, 2009; Flavell, 1979; Nelson & Narens, 1990). Although this definition as it stands might encompass thoughts about the thoughts of others (otherwise known as "mentalizing" or "theory of mind"), the term is generally understood as restricted to thoughts about one's own thoughts, as well as thoughts about one's own mental states and processes more generally. That is how it will be used here, too—at least initially. (The definition will be broadened herein to include other kinds of representation of one's own mental states, in addition to thought-like ones.) Note that both metacognition and mentalizing are meta-representational—they involve representations of one's own or others' representational states.

Metacognition is believed to be a vitally important capacity, with implications for education, emotional regulation, and self-awareness generally (Dunlosky & Metcalfe, 2009; Fleming, 2021; Gross, 2015). It has been heavily investigated in humans since Flavell's (1979) groundbreaking work, for the most part employing explicitly-expressed (verbal or numerical-scale) metacognitive judgments. In a paradigmatic explicit task used with humans, for example, participants perform some task and then report judgments of their confidence in their performance. The closer the correspondence between actual (objectively-measured) performance and reported confidence, the better the participant's metacognition is said to be.

Comparative psychologists, too, have sought evidence of metacognitive abilities in other creatures, employing designs from which, it is alleged, metacognitive awareness can be inferred. There have been two main paradigms used in the comparative literature. One has involved tests of uncertainty monitoring, in which animals have been shown to selectively opt out of difficult test trials without a penalty, or to adaptively accept or decline a high-stakes gamble on the correctness of their performance (Smith, Shields, & Washburn, 2003, Smith, Beran, Redford, & Washburn, 2006; Kornell, Son, & Terrace, 2007; Couchman, Coutinho, Beran, & Smith, 2010). The other paradigm has been designed to show an animal's awareness of its own knowledge or ignorance (Hampton, 2001, 2005; Rosati & Santos, 2016). Claims about the extent to which performance on these tasks reveal genuine, human-like, metacognitive ability have been heavily contested, however (Carruthers, 2011; Carruthers & Williams, 2019; Jozefowiez, Staddon, & Cerutti, 2009a, 2009b; Le Pelley, 2012).

We will here propose a three-way division among the mental

* Corresponding author.
E-mail addresses: pcarruth@umd.edu (P. Carruthers), D.M.Williams@kent.ac.uk (D.M. Williams).

https://doi.org/10.1016/j.cognition.2022.105117
Received 30 September 2021; Received in revised form 25 March 2022; Accepted 31 March 2022

processes involved in such studies. Some, we suggest, don't involve metacognition at all, but are purely first-order in nature, involving interactions among mental states, perhaps, but no explicit meta-representations of those states. We will argue, in particular, that nonverbal uncertainty-monitoring paradigms fall into this category, whether employed with humans or animals. But then there are two separate kinds of process that do genuinely involve self-directed meta-representations, we will suggest. One is "model-based" metacognition, in which the representations in question form components of an organized theory-like structure of such representations, of the kind investigated and described in classic studies of human metacognition. The other is "model-free," in which singular signals with meta-representational contents guide some down-stream first-order process or decision in the absence of anything remotely resembling a theory of one's own mind, and without being embedded among a set of concept-like representations of one's own mental states or processes.[1] We will suggest that some forms of model-free metacognition are present in other animals, but that evidence of model-based metacognition is currently lacking.

The language of "model-based" versus "model-free" metacognition is introduced by analogy with model-based versus model-free learning and decision making (Dayan & Berridge, 2014; Dickinson & Balleine, 1994, 2002; Gläscher, Daw, Dayan, & O'Doherty, 2010). In model-based decision making a representation of the causal structure of the environment gets built, and actions are selected by searching through and evaluating the options permitted by the model. In model-free decision making, in contrast, actions acquire a sort of "cached value" through evaluative learning, and are selected automatically and independently of changes in the causal structure of the environment. In model-based metacognition, likewise, there is some sort of causal model of aspects of one's own mind that is used to guide metacognitive interventions, whereas model-free metacognition would operate automatically and without needing guidance from any such set of representations.

The differences between models and theories have been extensively discussed by philosophers of science (Giere, 1988; Godfrey-Smith, 2006; Nersessian, 2002). Models are often thought to employ structural or map-like representations whereas theories are framed in terms of causal generalizations (Butlin, 2021). Such differences practically disappear when the theories in question are implicit ones, however, with generalizations being replaced by associative connections or inference rules linking explicit representations (Gopnik & Glymour, 2002). For our purposes, what matters more is that both models and theories are always understood to be structured entities, comprising multiple nodes or concept-like representations, together with the connections between them.

Model-*free* forms of explicit metacognition, in contrast, would comprise lone representational signals (whether concept-like or analog-magnitude) that have mental states of the self among their truth-conditions or correctness-conditions.[2] These signals would serve to guide some down-stream cognitive process independently of any other metacognitive representation (and so independently of any meta-cognitive model). The two cases that will concern us in due course are stand-alone metacognitive belief-like representations with the content, *[that is] unknown*, which serve to drive curiosity and interest; and meta-acognitive analog-magnitude signals representing degrees of executive-system engagement, whose appraisal produces feelings of cognitive effort.

Before proceeding further, it may be helpful to pause briefly to explain how the terms "implicit" and "explicit" are being used throughout. (See Table 1 for summary definitions of a longer list of terms that may be unfamiliar to some readers.) By explicit representation and meta-representation, we do not mean (meta-)representations that are conscious. (While the explicit / implicit contrast is sometimes intended to line up with the conscious / unconscious one, that is not how it will be employed here.) Rather, we mean that they involve some form of (meta-)representational symbolic structure or signal, as opposed to being built tacitly into the processing rules or procedures employed. Relatedly, an explicit *task* is one whose output measure is an explicit representation of some sort (e.g. verbal or numerical-scale), whereas an implicit task is one whose output measure is not a representation (but rather, e.g., a choice among options, or a response-time). Note that an implicit *task* may nevertheless be enabled and explained by explicit meta-representations (that is to say, by explicit metacognition, whether model-based or model-free). Whether or not, and when, that is so is the topic of the present article.

Likewise, by an explicit theory or model, we mean one whose generalizations are explicitly represented in symbolic structures. To illustrate, the information that seeing leads to knowing is *explicitly* represented by the mentalizing system (albeit unconsciously) if the inference from, "John sees that P" to, "John knows that P" is mediated by consulting the major premise, "Seeing leads to knowing." In contrast, the information is *implicitly* represented if "John sees that P" leads directly to "John knows that P" through a built-in domain-specific inference rule having the form: "X sees that P → X knows that P." Notice that seeing and knowing, in contrast, are explicitly represented either way, by the concepts SEE and KNOW respectively. So an implicit *theory* can nevertheless govern explicit *representations*.

Some in the field have sought to characterize forms of metacognition that are wholly implicit. This is so-called *procedural* metacognition (Dokic, 2012; Proust, 2014). This view postulates no symbols or explicit signals with meta-representational contents. Rather, a cognitive process that monitors and modulates the outputs of another gets described as procedurally metacognitive, even if it fails to rely on any model-based, or theory-like, understanding of the self, and even if no meta-representations of mental states or processes are involved at all. Still the procedures in question are said to qualify as metacognitive in nature, because they function to modulate and control ongoing cognitive activity.

This use of the term "metacognition" comes too cheap to be worth the name, however (Carruthers, 2017). For example, using this definition, the operations of the bottom-up attentional salience system (Corbetta, Patel, & Shulman, 2008) would qualify as implicitly metacognitive, because of its role in monitoring the significance of unattended (unconscious) representations and in modulating the direction of top-down attention, thereby influencing the contents of working memory. Yet, no attention researcher has ever claimed that such a bottom-up system is meta-representational; nor does it need to be so, in

[1] We use the phrase "concept-like," here, because some philosophers object to ascribing concepts to animals at all. This is on the grounds that the states of animals fail to satisfy the "Generality Constraint" on concept-possession (Bermúdez, 2003; Camp, 2004; Evans, 1982). (The Generality Constraint maintains that if a creature is capable of employing any concepts at all, then it must be capable of freely recombining any of the concepts that is has with any of its other concepts of appropriate adicity. So if it has concepts *F* and G, and singular concepts *a* and *b*, then it must be capable of thinking each of the permissible combinations: *Fa*, *Fb*, *Ga*, and *Gb*.) While this view is arguably not defensible (Carruthers, 2009a), that debate is not relevant here. No substantive questions are begged by us opting to speak in terms of *concept-like* and *thought-like* states in animals instead.

[2] Beliefs and belief-like states typically have truth-conditions; in contrast, low-level perceptual states and analog-magnitude representations generally have correctness-conditions (Beck, 2018). Representations of approximate numerosity, for example, can be more or less *accurate*, without being categorically true or false. See the summary definitions in Table 1.

Table 1
Terms employed in this article.

Analog-magnitude representation	Representation of a quantity or magnitude by means of a continuous (as opposed to discrete) representational vehicle. Commonly employed in perceptual systems, but also in representations of approximate number etc.
Competitive accumulator	Neural representations that build their activity over time in mutual competition with others
Conceptual representation	An explicit signal or symbol that categorically picks out some component or aspect of a domain. When centrally available, concepts are the building blocks of thought
Correctness condition	The condition under which an analog-magnitude or nonconceptual representation is accurate or correct. This is generally a matter of degree rather than all-or-nothing
Explicit mental representation	A neural signal or symbolic structure that has, or contributes to, the truth-condition or correctness-condition of a mental state
Explicit task	Task whose behavioral output measure is an explicit (e.g. verbal or numerical-scale) representation
Implicit mental representation	A mental process or inferential disposition that contributes to the truth-condition or correctness-condition of a mental state, but does so without explicit representation
Implicit task	Task whose behavioral output measure is not an explicit representation
Mentalizing	The process of attributing mental states to other creatures (and also, on some views, to oneself). Often called "theory of mind"
Metacognition	The process of attributing mental states to oneself, or representing mental processes in oneself (whether conceptually or non-conceptually)
Meta-representation	Any representation that represents another representation, whether in language or in the mind. Mentalizing and metacognition both traffic in meta-representations
Model-based	A process that operates via a structured model or theory of a domain, always containing more than a single component. One signature of a model-based process is flexibility in the face of changes in the domain represented.
Model-free	A process is model-free when it is not model-based (generally utilizing just a single representational signal). It operates inflexibly, and is not sensitive to changes in the structure of the domain represented
Nonconceptual representation	A nonconceptual representation is one that does not "chunk" or categorize the domain represented. Generally an analog-magnitude representation
Procedural metacognition	A mental process or procedure that adaptively alters one's own mind in some way, but without employing any meta-representations
Truth condition	The condition under which an assertion or thought is true or correct. Truth is generally all-or-nothing, although also allowing for borderline / indeterminate cases

order to perform its role in guiding attention.

Our topic, in contrast, is *explicit* (symbolically represented) meta-representational forms of metacognition—that is, forms of metacognition that involve symbolic signals or structures of some sort with self-directed meta-representational contents. In effect, our topic is whether, and when, there are explicit representational signals in the mind that have some of the mental states or processes of the agent among their truth-conditions or correctness-conditions—that is to say, among the represented *contents* of those signals—and whether those signals are model-free or model-based. More specifically, our question is whether there are any explicit metacognitive signals that are *not* part of a flexible network of such signals, but rather perform their role singly. Such signals would be designed to do their work in a model-free manner, and are likely to be of ancient provenance.

How can these distinctions be tested for empirically? Sections 2 and 3 will illustrate one way in which the presence or absence of model-based metacognition can be tested for in nonverbal creatures. (Standard tests of metacognition among humans using verbal reports, numerical-scale judgments, and other communicative measures are obviously tests of model-based metacognition. Moreover, as we will see in Section 2, they have always been understood as such.) Then Sections 4 and 5 will examine two possible instances of model-free metacognition, rendering a negative verdict in each case. This will depend on our understanding of the detailed computational processes underlying them, and their contents. But then Section 6 will argue that curiosity and other forms of search behavior motivated by ignorance or memory failure do genuinely demonstrate the presence of model-free metacognition. Section 7 will make a similar argument for some of the signals underlying decisions to deploy cognitive effort in executively demanding tasks. Finally, Section 8 will conclude and look ahead to potential empirical testing.

Before embarking on those discussions, however, it is worth asking how model-based metacognition can be distinguished from model-free metacognition when the models in question are implicit ones. For it might be said that *any* meta-representational signal will be embedded in computations involving other representations of *some* sort. And then the question is why those doesn't count as an implicit model of the domain. We stressed above, however, that model-based processes always comprise multiple explicit representations, even when the model itself is implicit in a set of associations or inference-rules linking them together. And model-based metacognition, in particular, must then involve multiple types of meta-representational signal, rather than just one. Model-free metacognition, in contrast, would involve processing over just a single such signal, and the representations that interact with or are caused by that signal will be first-order ones. For example, if it is signals with the content, *unknown*, that issue in curiosity, then they will activate (more-or-less strongly) non-metacognitive motor plans for such things as approaching closer, sniffing the target object, and so on.

2. Model-based metacognition in humans and other animals

For the most part psychologists who study human metacognition have focused on the determinants and accuracy of a variety of explicitly-expressed metacognitive judgments, indicated verbally or on a numerical scale of some sort (Dunlosky & Metcalfe, 2009). These include judgments of learning and ease of learning, judgments of confidence, expressions of feeling-of-knowing and tip-of-the-tongue states, and judgments about the sources of one's own knowledge.

Thus understood, metacognitive judgments are always meta-representational—they involve explicit representations of mental states or processes in oneself. But they also rely on an implicit model or "theory" of the operations of one's own mind. According to the standard theoretical approach to classifying and characterizing metacognitive processes in humans, there is a meta-level that monitors, represents, and controls the processes within object-level cognitive systems; and it has always been an important aspect of such accounts that the meta-level contains a meta-*model* of the object-level (Nelson & Narens, 1990, Fig. 1, and p.126, Principle 2: "The meta-level contains a dynamic model … of the object-level"). It is the meta-model that is used to guide interventions to alter the course of one's own cognitive processes—to improve one's learning, say, or when deciding how much reliance to place on a previous judgment. So the kinds of human metacognition standardly investigated by psychologists can be described as *model-based*.

Many in the field have thought that the implicit model of one's own mind that guides metacognitive control processes is the same as, or is an extension of, the implicit model or "theory" that guides our predictions and explanations of the mental states and behavior of other people (Carruthers, 2009b, 2011; Frith & Happé, 1999; Gopnik, 1993; Perner & Ruffman, 1995; Wellman, Cross, & Watson, 2001; Williams & Happé, 2009). Those espousing such a view have generally held that human mentalizing abilities have priority over metacognitive ones, in both phylogeny and ontogeny. The former claim is grounded in theories that emphasize the importance of meta-representation for "Machiavellian

intelligence" and social coordination more generally (Byrne & Whiten, 1988; Seed & Tomasello, 2010; Whiten & Byrne, 1997). The latter claim receives some support from evidence that capacities to represent the mental states of other agents are present in human infancy, even among infants as young as six months of age (Baillargeon, Scott, & Bian, 2016; Hyde, Simon, Ting, & Nokolaeva, 2018; Scott & Baillargeon, 2017; Southgate & Vernetti, 2014),[3] whereas there is no evidence of metacognitive capacities in the first few years of life.

Alternatively, some have claimed that awareness of one's own mental states is more basic, with capacities for attributing mentality to other agents depending on metacognition together with simulative and imaginative abilities (Gallese & Goldman, 1998; Goldman, 2006). Such a view would be supported by finding model-based metacognitive abilities in creatures that are incapable of equivalent forms of mentalizing (at least, if the reports of such findings were to hold up; see Section 3).[4]

Those who study metacognitive processes in animals often cite the Nelson & Narens model with approval (e.g., Smith et al., 2003, Smith et al., 2006; Couchman et al., 2010), so one can assume that when the animals in question are claimed to have metacognitive capacities, it is generally some version of such a model-based architecture that they are thought to possess.[5] It need not be part of the view, of course, that the mind-model in question is anything like as rich or as structured as the human mentalizing system. But if the metacognitive abilities attributed to these animals are to support claims of nascent self-awareness in these creatures (Couchman, Coutinho, Beran, & Smith, 2009), or to warrant titles such as, "Rhesus monkeys know when they remember" (Hampton, 2001), then they must implicate a mind-model of *some* sort. Moreover, it is often the case that positive findings with nonhuman animals are described as evidence of early forms of the kinds of metacognition found in humans (e.g., Rosati & Santos, 2016). In contrast, Section 3 will review recent data from two sets of experiments suggesting that there is little or nothing in common between many of the sorts of behavioral "metacognitive" tasks employed with animals and the explicitly-expressed judgments investigated in human metacognition research.

3. Testing the tasks used to measure model-based metacognition in monkeys

There have been extensive debates about the phylogenetic origins of self-awareness. Some experimenters have claimed, for example, that the success of monkeys or other animals in so-called *uncertainty-monitoring* tasks manifests at least a simple form of metacognitive awareness of their own mental states—either as such, or in a similar enough manner that the representations in question are preadapted to become components of full-blown self-awareness in humans (Smith et al., 2003, Smith, Couchman, & Beran, 2014; Kornell et al., 2007; Couchman et al., 2009, 2010). In all such tasks the animals have to make primary discrimination of some sort, of varying difficulty. But in some paradigms the animals are given the opportunity to opt out of trials where they are uncertain, moving on to the next trial without reward or penalty (Smith et al., 2003). Another paradigm requires the animal to take either a high-stakes or low-stakes gamble on the correctness of its initial choice (Kornell et al., 2007).

We assume then (in light of the discussion in Section 2), that such experimenters are claiming to discover at least simple forms of model-based metacognition in monkeys. Some critics have charged that the findings can be explained away in associative terms (Le Pelley, 2012). Others have appealed instead to first-order estimations of risk, or have claimed more generally that the epistemic emotions in question (uncertainty, in particular)—and in contrast with the explicit judgments in humans that those feelings can ground—are likewise first-order (non-metacognitive) in nature (Carruthers, 2017; Ritchie & Carruthers, 2012).

If the kinds of uncertainty-monitoring tasks conducted with monkeys are genuinely tapping into capacities for model-based metacognition (albeit nascent, and simplified in comparison with human forms of metacognition) then a number of predictions can be made. These were tested in two recent sets of experiments (Nicholson, Williams, Grainger, Lind, & Carruthers, 2019, 2021).[6] The predictions are as follows.

First, such tasks should share at least some metacognitive resources with equivalent explicit (e.g. verbal) tasks that employ the same materials and same basic structure. In which case, performance in the two types of task should be significantly (but of course not completely) correlated in humans.

Second, recall that on either of the relevant accounts, metacognition and mentalizing are linked. (It is either the case that model-based metacognition involves self-directed mentalizing or else mentalizing is grounded in metacognitive self-awareness.) As a result, we should expect performance in both explicit and implicit uncertainty-monitoring tasks to be correlated with mentalizing abilities in humans.[7]

Third, no matter what the direction of the relationship is between metacognition and mentalizing, we should predict that a concurrent mentalizing task will disrupt both explicit and implicit metacognitive tasks in neurotypical people. Of course, if metacognition requires self-directed mentalizing, then taking up mentalizing resources through a secondary task should disrupt performance in a metacognitive one. But even if mentalizing depends instead on a combination of metacognition and mental simulation, a concurrent mentalizing task will nevertheless require metacognition as well as simulation. Hence concurrent mentalizing should still disrupt both explicit and implicit metacognitive task performance.

Fourth, depending on the direction of the relationship between metacognition and mentalizing, we can make predictions regarding the

[3] There are now well over 30 studies that provide evidence of false-belief understanding in infants and young children, using a variety of materials and methods, and coming out of a number of different labs (Scott & Baillargeon, 2017). Admittedly, there have recently been some failures to replicate individual findings (for examples: Dörrenberg, Rakoczy, & Liszkowski, 2018; Kammermeier & Paulus, 2018). But Baillargeon et al. (2018) point out the methodological weaknesses of many of these attempted replications, while also acknowledging that some methods (specifically anticipatory looking) might not be reliable. And in the meantime, new studies both replicating and extending previous findings continue to be published (Buttelmann & Kovács, 2019; Forgács et al., 2019; Király, Oláh, Csibra, & Kovács, 2018).

[4] There is a third possible view of the relationship between metacognition and mentalizing. Nichols and Stich (2003) argue that they are independent of one another—sharing no resources, and with the possibility that each can be damaged or absent while the other is fully intact. This third view has received little empirical support, and will not be considered further in the current work. (See Carruthers, 2011, for an extended critique.)

[5] In fact, there is some confusion in the literature on this point. Sometimes researchers in the field are focused on establishing that the animals in their experiments are relying on more than merely associative processes, claiming, in contrast, that their responses are executively controlled (hence being "metacognitive" only in the weak sense of being "above" other cognitive processes in a hierarchy of control, rather than involving meta-representations). This is then purely procedural "metacognition," of the sort discussed briefly (and rejected as not worthy of the name) in Section 1. See Carruthers (2014) for discussion, commenting on Smith et al. (2014).

[6] In addition, (alongside the predictions discriminating between meta-representational and first-order accounts of the "uncertainty monitoring" tasks conducted with monkeys described here), both experiments involved a number of pre-registered predictions that were designed to discriminate between the theory that metacognition is self-directed mentalizing, on the one hand, and an account of mentalizing as other-simulating metacognition, on the other. The results confirmed our prediction that is it the third-person mentalizing system that is at least partially responsible for self-directed metacognition.

[7] Recall that we are setting aside the view that metacognition and mentalizing are independent of one another in humans. See footnote #4.

performance of people with autism spectrum disorder (ASD), whose mentalizing abilities are compromised.[8]

This fourth set of predictions will take a little explaining.

On the one hand, suppose that metacognition involves self-directed mentalizing (as we have previously suggested is the case; Carruthers, 2009b, 2011; Williams, 2010). In that case, if both explicit and implicit uncertainty-monitoring tasks involve model-based metacognition, then one would expect performance in both sorts of tasks to correlate with mentalizing abilities in humans. This is because both types of task require meta-representation of one's own uncertainty. Hence one would also expect performance in both kinds of task to be poorer in people with ASD, who have well-documented difficulties with mentalizing.

Then suppose, on the other hand, that metacognitive abilities are more basic, and underlie mentalizing ones when combined with capacities for other-directed imagination or simulation, as some have suggested (e.g., Goldman, 2006). There are then two possibilities for the component that is compromised in ASD (either metacognition or imagination—or both, of course, but that would then issue in the same set of predictions already outlined above). If what is compromised in ASD is the metacognitive component of mentalizing, then again we should expect performance in both explicit and implicit metacognitive tasks to be diminished in autistic people. (Note, though, that we know of no one who has actually suggested that it is the metacognitive component of mentalizing that is diminished in ASD.) But conversely, if what is compromised in ASD is the simulative component of mentalizing (as simulation theorists typically claim; Goldman, 2006), then we should expect that performance in both explicit and implicit metacognitive tasks should be *intact* in ASD. Hence, either way, if these implicit tasks are genuinely metacognitive in nature, then we should expect, *either* that performance in both kinds of task will be compromised in ASD, *or* that performance in both should be intact in ASD.

The bottom line for the third and fourth predictions: no matter what the relative priority between metacognition and mentalizing is, if implicit uncertainty-monitoring tasks of the sort conducted with monkeys tap into any form of model-based metacognition, then we should expect a concurrent mentalizing task to disrupt performance in *both* the explicit *and* the implicit tasks in neurotypical people. And whatever the relative priority of metacognition and mentalizing, we should expect *either* that both explicit and implicit task-performance are intact in ASD, *or* that both should be damaged together. What we should *not* predict is that performance in implicit tasks of the sort conducted with monkeys would be intact in autistic people, or among neurotypical people when completing the tasks alongside a secondary mentalizing task, whereas performance in otherwise-matching explicit ones is deficient and/or disrupted.

Nicholson et al. (2019, 2021) tested all of these predictions using two of the types of uncertainty-monitoring task that have been employed with monkeys. But they had human adults and children perform the tasks both implicitly (making a behavioral choice depending on their degree of certainty, just as the monkeys do) and explicitly (making a verbally-expressed judgment of confidence). In some of their experiments, the performance of autistic adults or children was compared with the performance of matched control participants. In other of their experiments, neurotypical participants performed either the implicit or explicit version of the task while completing a concurrent metalizing task. The performance of these participants in a dual-task condition was compared to the performance of participants who completed the metacognitive tasks alone in a single-task condition, as well as in a variety of other dual-task conditions used as controls.

In both sets of experiments, participants first had to make perceptual discriminations of varying difficulty (this was the object-level task); and in both cases, in the explicit-judgment condition, participants recorded their judgment of confidence once they had done so. But Nicholson et al. (2019) used the opt-out method employed by Smith et al. (2003) for the implicit condition—participants had the option of skipping the test and moving directly to the next trial without gain or loss. Nicholson et al. (2021), in contrast, used the implicit gambling method employed with monkeys by Kornell et al. (2007). Having made a discrimination, participants had the option of selecting an abstract "high risk" symbol that would give a large payoff if the initial response was correct, but a large loss if it was incorrect; or a "low risk" symbol that had small positive or negative payoffs. Great care was taken not to signal to participants that the implicit tasks could be approached by first making an explicit judgment of certainty, and also to insure a matching incentive structure for both explicit and implicit tasks.

The findings from each of these two sets of studies are inconsistent with the view that implicit as well as explicit tasks tap into model-based forms of metacognition. *First*, performance in the two types of task was uncorrelated, suggesting that cognitive resources are not shared.[9] *Second*, performance on the explicit tasks, but not the implicit tasks, was correlated with mentalizing abilities, suggesting that only the former type of task depends on meta-representation. *Third*, the secondary mentalizing task selectively interfered with explicit-task performance, while leaving intact performance in tasks of the sort conducted with monkeys.[10] And *fourth*, performance was impaired among ASD participants in the explicit verbal task, but not in the implicit task.[11]

The evidence suggests, then, that the kinds of implicit uncertainty-monitoring tasks conducted with nonhuman animals have nothing in common with the forms of explicit (often verbal) metacognitive tasks routinely employed with human subjects. This suggests that the implicit tasks aren't tapping into model-based metacognition, in the way that standard forms of task plainly are. We cannot conclude from this, of course, that animals lack all capacity for model-based metacognition—nor even that Macaque monkeys do. For there are other forms of task that have not been discussed here (Hampton, 2001, 2005; Beran et al., 2015; see Carruthers, 2017, and Carruthers & Williams, 2019, for discussion and critique). But even if this strong conclusion were accepted, that would leave open that some of the implicit tasks

[8] Mentalizing abilities are not completely lacking in ASD, of course; and it remains controversial how fundamental the mentalizing deficit is within the disorder as a whole. But it is well established that mentalizing abilities are *diminished* among people with ASD in comparison with neurotypical controls (Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998; Brunsdon & Happé, 2014; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Jones et al., 2018). This is all that we require for our purposes here.

[9] Although this is a null result, it is one that we specifically predicted. Moreover, we subsequently conducted a Bayesian analysis of the data from Nicholson et al. (2019) (n = 43 in each reanalysis). This enables one to test whether the data provide positive evidence for the lack of correlation. Our finding has a BF^{10} of 0.258, which means that the data are 3.88 times more likely to support the null hypothesis than the alternative hypothesis. Likewise, reanalysis of the data using Bayesian techniques confirmed that the association between explicitly-expressed metacognitive accuracy and mentalizing task performance was not only statistically significant (r = 0.38, p = .01), but associated with a BF^{10} of 4.28. In contrast, the association between nonverbal task performance and mentalizing was non-significant (r = 0.12, p = .43) and associated with a BF^{10} of 0.22. In other words, the data were 4.28 times more likely to support the alternative than the null when it came to the relation between mentalizing and explicit metacognition, but 4.54 times more likely to support the null than the alternative when it came to the relation between mentalizing and adaptive opt-out behavior.

[10] Importantly, selective interference of the concurrent mentalizing task on the verbal metacognitive task was not merely because the two tasks shared language demands; Nicholson et al. (2021) employed a number of other verbal secondary tasks that did not tap mentalizing and none impaired verbal metacognitive performance.

[11] Likewise, the participants with ASD were matched closely to control participants for age and verbal ability, so selectively impaired performance on the verbal metacognitive task was not merely because of the language demands of the task.

employed with animals might manifest model-*free* forms of metacognition. We will return to this suggestion in Section 5 (with respect to uncertainty monitoring, but arguing for a negative conclusion) and again in Section 6 (discussing curiosity, memory monitoring, and information search, this time arguing for a positive one). Before that, in Section 4, we consider the one alleged instance of model-free metacognition that has hitherto been defended in print. This will provide us with a template for our investigation going forward.

4. Are error-signals metacognitive?

A claim of this sort was first proposed by Shea (2014), who argues that the error-signals involved in affective / reward-based learning are meta-representational. This is on the grounds that they represent the difference between an *expected* and an *experienced* reward. Note that affective learning involving such signals is extremely widespread in the animal kingdom, and is found even among snails (Kobayashi et al., 1998) and ants (Cammaerts, 2004), as well as among all birds and mammals. Yet no one would want to claim that ants and snails exhibit nascent forms of self-awareness as a result. Nor, we suggest, would anyone ever publish a paper with a title like, "Ants sometimes know that their expectations of value aren't met."[12]

Shea (2014) argues at length that reward-prediction error signals are nonconceptual (and model-free, although he doesn't use the term) forms of explicit metacognition. They represent the difference between a predicted and an experienced reward, and thereby serve to update the agent's representation of the reward-value of the entity or action in question. But these error signals are mostly buried deep in subcortical regions of the brain, and are common to all creatures capable of evaluative learning, including many invertebrates. No one would think of them as involving a nascent form of self-awareness. Nevertheless, in Shea's telling, they are meta-representational in content, representing the magnitude of the difference between a *predicted* and an *experienced* reward.

One might think that even the language used by theorists in the field—*error* signal—suggests a meta-representational content, implying that what is signaled is that a representation (the content of a prediction) is erroneous or mistaken. It is not obvious that the term employed is anything more than a theorist's external gloss, however. We as theorists can see, of course, that an expectation has been formed and then disconfirmed—that the expectation was erroneous. But it doesn't follow from this that the content of the error signal itself represents that a representation is mistaken. Moreover, the standard way of stating the content of the error signal—that it represents the difference between an expected value and an experienced value—admits of two different readings, corresponding to differences in the scope of "represents." It can either mean (as Shea suggests): "The error signal represents: [the difference between an expected value *m* and an experienced value *n*]." Or it can mean: "Concerning an expected value *m* and an experienced value *n*, the error signal represents: [the difference between *m* and *n*]." On the latter reading, an error signal represents the difference between two (represented) *values*, not two *representations* of value.

Since the error signal is caused by a mis-match between an expected and an experienced value, it does carry *information* about each of these mental events, of course. But as has long been recognized in philosophy, carrying of information is not sufficient for representation (Dretske, 1986, 1988; Fodor, 1990; Millikan, 1989). The state of perceiving a cat, for example, carries information about a pattern of stimulation on the retina and patterns of light transmitted through the air, as well as that there was a historical cat-mating that led to the existence of that cat. But it *represents* none of those things. And closer to home, many shifts of attention carry information that a prior appraisal of the relevance of a stimulus or memory has occurred (Corbetta et al., 2008). But of course, shifts of attention, although often caused by appraisal processes, don't *meta-represent* those processes.

In more recent work, Shea (2018) has reviewed and synthesized previous theories of representational content, drawing especially on the role that appeals to representation play in cognitive science. He argues that what makes it the case that something is a representation is that it plays a computational role in some cognitive process. And what fixes the content (or correctness condition) for a representation from among all the information that it carries, is what causally *stabilized* the role of that representation in the computations that it enters into, either through natural selection, or through learning, or through its contribution to individual survival. The content of the representation is the carried-information that we need to appeal to in explaining the role that the representation plays in determining the behavior of the organism. We assume that this account is at least approximately correct, although the details won't matter for our purposes.[13]

Now, error-signals are analog-magnitude representations that can be either positive or negative, depending on whether the experienced value is greater or lesser than expected (and by how much). Let *plus-d* be a measure of how much greater the experienced value *n* is than the expected value *m*. Then, given the way in which error-signals like *plus-d* are reliably caused, they do carry the information that an expectation of value had previously been formed while a greater value was subsequently experienced. So they do carry metacognitive *information*. But they also carry information about the difference between the stored value for instances of a kind of thing (which might be an initial innate setting, or which may have been learned previously from encounters with the environment) and the value of the current item. Hence we can ask (using Shea's own semantic framework) which of these sorts of information played a causal role in stabilizing the evaluative-learning system with the properties and causal role that it now has.

The answer is obvious once the question is clearly posed. What matters from the perspective of evolution is accurately tracking and updating the adaptive value of items and actions in the environment. So what matters is generating an accurate representation of the magnitude of the difference in value between the current item or action and the values possessed by members of the kind that the organism (or its ancestors) has previously encountered. In effect, the correctness condition for *plus-d* concerns how much better this *X* is than *Xs* in general. This is a first-order representation, lacking any metacognitive content.

There is more to be said on this topic, of course. Carruthers (2021) elaborates on these criticisms of Shea's (2014) interpretation of

[12] Perhaps for this reason, Shea (2014) describes these signals as meta-representational rather than metacognitive in nature—maybe he thinks they are too far distant from standard forms of metacognition to be described as such. Since our interest, in contrast, is in showing that there are two *kinds* of metacognition (model-based and model-free—note that this terminology is introduced here for the first time), we propose to present his arguments in our own terms.

[13] Note, however, that Shea (2018) actually defends what he calls "varitel semantics," which includes *two* basic kinds of representing-relation. One is informational, with an internal symbol causally co-varying (in the right circumstances and in the right way) with the represented property or thing. The other is a form of structural mapping, with the relations among a set of internal symbols mirroring the relations among a set of external entities in a map-like manner. It is the first of these sorts of representing relation that is relevant here. This is because error signals are singular in occurrence, rather than doing their work via the relations they stand in to a set of similar signals. Note, too, that Shea's (2018) framework is employed here for concreteness, and because it is the best-developed recent theory of the role of content-assignments in cognitive science. Although Shea's account has a significant teleological flavor ("the content is the information carried that explains how the role of the symbol-type in question got stabilized"), if we were to switch to an account like that provided by Rupert (2018), this would have no impact any of our conclusions. Rupert emphasizes, instead, the role of the information carried in explaining down-stream uses, and ultimately the organism's current behavior.

evaluative error-signals in some detail. What matters at this point is that we have a template for investigating whether there are forms of model-free metacognition in other systems, and underlying other cognitive phenomena. What we need to be looking for, in each case, are explicit signals of some sort that turn out to have other mental states or processes among their correctness conditions when we apply some version of Shea's (2018) semantic framework. This is the strategy to be pursued for the remainder of our discussion, beginning with a second look at the forms of "uncertainty monitoring" behavior displayed by monkeys. Although there is good reason to deny that such behavior manifests any form of model-based metacognition (as we saw in Section 3), that leaves open the possibility that there might be metacognitive signals underlying it that are model-free.

5. Certainty and uncertainty revisited

Many theorists think that when humans make explicitly-expressed metacognitive judgments—about their own learning, memory, accuracy, certainty, and so on—those judgments are always heuristic-based rather than directly introspective (Dunlosky & Metcalfe, 2009; Koriat & Goldsmith, 1996).[14] Lacking access to memories that are currently inaccessible (in feelings of knowing), or lacking access to the processes that constitute learning (in judgments of learning), people are forced to fall back on indirect heuristic cues. Other theorists think that in the domain of certainty and uncertainty, in particular, the probability estimates that underlie first-order choice can be made available to inform explicit metacognitive judgments as well (Bahrami et al., 2010). We allow that this is possible. Since verbal expressions of certainty are themselves a form of behavior, it is possible that they should likewise be influenced by the probability judgments that underly choice behavior. But we are doubtful. For the data reviewed in Section 3 found no overlap between implicit tasks—which we interpret as probability-based—and explicit metacognitive judgments. Moreover, even if probability estimates could inform explicit forms of metacognition, this would not show that implicit tasks are themselves metacognitive, of course (in either a model-based or a model-free sense).

It seems that when humans make explicit (verbal) judgments of certainty or uncertainty there are a number of heuristic cues that influence them. One is reaction-time (Kiani, Corthell, & Shadlen, 2014). If one is slower than normal to reach a decision (about which of two lines is longer, for example), then that is an indication that one is uncertain of the correct answer. Thus interventions that artificially inflate reaction-times without impacting accuracy lead to lower confidence judgments. Another cue that people frequently rely on when making explicit judgments of certainty is fluency versus disfluency of processing (Rhodes & Castel, 2009; Kornell, Rhodes, Castel, & Tauber, 2011). Fluent processing produces positive affect, and disfluent processing is negative (Carr, Rotteveel, & Winkielman, 2016; Casasanto & Chrysikou, 2011). Roughly speaking the heuristic is: if making a first-order judgment makes you feel bad, then say you are uncertain of its correctness. Hence manipulations that impact fluency without altering accuracy (like increasing or decreasing the point-size of words to be studied, or using familiar rather than unfamiliar materials in a reasoning task) increases or decreases people's confidence (Ackerman & Thompson, 2017).

It seems that people will say that a perceptual judgment they have just made was an easy one, and likely to be correct, provided they made that judgment swiftly, and/or felt good when doing it. Conversely, the cues people use when reporting that a judgment is likely to be incorrect (because hard) are that it took a while for them to reach a decision and/or they felt bad while doing it. The first of these heuristic cues (reaction time) is not in any sense meta-representational. Representations of reaction-times are first-order analog-magnitude representations of the temporal intervals between stimulus presentations and actions (Odic et al., 2016). As we will see in Section 7, in contrast, there is some reason to think that signals of cognitive effort of the sort involved in disfluent processing are (at least in other contexts) model-free metacognitive ones. So it appears that when humans make explicit (model-based) judgments of certainty their judgments may depend in part on metacognitive signals (signals that represent executive-system engagement). It is another matter, however, to claim that these same signals underlie the performance of non-human animals in uncertainty tasks (as well as humans in implicit versions of the same tasks). Indeed, if they did, it would be puzzling that there should be no overlap in performance between explicit and implicit versions of the same tasks, as we saw in Section 3. Moreover, as we will see shortly, performance in implicit uncertainty tasks seems best explained in purely first-order (non-metacognitive) terms. It thus turns out that it is one thing to *use* one's uncertainty in an on-line manner, and quite another thing to make a heuristic-based judgment *about* one's uncertainty.

Investigations and formal modeling of perceptual uncertainty in humans (using the sorts of materials that might be employed with nonhuman animals, such as judging which of two lines is longer) have issued in a debate between those who think that confidence-formation is inherent in the decision process itself (van den Berg, Anandalingam, et al. (2016); Khalvati, Kiani, & Rao, 2021), and those who find that it involves an additional second stage of processing that is second-order or hierarchical in nature (Fleming & Daw, 2017; Maniscalco & Lau, 2016). We note, however, that the latter set of studies employ explicit judgments of confidence, whereas the former do not. This maps nicely onto our suggestion that explicit metacognitive judgments express model-based metacognition, whereas implicit tasks of the sort discussed in Section 3 may not be metacognitive at all. So different computational models may be required to explain the data because the kinds of processing underlying the two sorts of task are distinct. That is what we will suggest here.

Only a few studies have not only used stimuli, but also response measures, of a sort that could be employed with animals.[15] But van den Berg, Anandalingam, et al. (2016) used a procedure with adult humans very much like the gambling task employed with monkeys by Kornell et al. (2007) and with humans by Nicholson et al. (2021), except that confidence was measured at the same time as the primary discrimination judgment. Participants judged the direction of motion of a random dot stimulus on each trial, indicating whether the dominant motion was leftwards or rightwards. They made their choice by moving a cursor to one of four positions on the screen, using a joystick beneath the table. The two left-most positions indicated leftward motion, and the two right-most positions indicated rightward motion. But the upper and lower positions on each side came with different payoffs in terms of

[14] Note that the term "heuristic" here is to be understood broadly, to contrast just with direct or introspective access to the mental process in question. It is not intended to exclude views that model people's use of heuristic cues in terms of boundedly rational algorithms, or within a Bayesian framework.

[15] In addition to the studies described below, Miyamoto et al. (2021) recently employed an implicit task with humans of a sort that might also be conducted with monkeys, claiming to find evidence of metacognition. Participants had to choose on each trial between two stimuli. One was a random dot motion stimulus, which if chosen would lead to a second stimulus with the same proportions of directed motion, where they would be required to judge the direction (for a fixed maximum reward if correct). The alternative stimulus just contained a number of dots in coherent motion, which indicated a likelihood of reward in proportion to the number of dots it contained. The authors interpret their task as contrasting metacognitive judgments with judgments of worldly probability. We are doubtful. Rather, at the initial choice stage participants had to make a decision between two probabilities: the probability of leftward motion (say) in the random dot motion stimulus versus the probability of reward in the fixed-motion case (estimated on the basis of numerosity). It is one thing to make a first-order judgment of probability, and quite another to make a metacognitive judgment of certainty. Unfortunately, these are often conflated in the literature.

points won or lost. The upper position was high-risk, giving a 2-point gain for a correct choice, but a 3-point loss for an incorrect one; the lower position was low-risk, with a 1-point gain or a 1-point loss. (Participants saw a running tally of their point scores throughout.) In effect, then, the upper positions expressed high confidence, but were not described to participants in these terms, and no explicit judgment of confidence needed to be made.

This setup enabled van den Berg, Anandalingam, et al. (2016) to collect data on accuracy, reaction times, confidence, and also changes of mind about the correct choice and/or about confidence. This is because the movement of the joystick could be precisely tracked, and on some trials participants started out toward one location (high-confidence about leftward motion, say) but switched mid-movement to end up at another (e.g. low-confidence about leftward motion). Participants were highly trained (just as monkeys in such experiments always are), and they completed over 9000 trials in total. The experimenters were then able to build a robust theoretical model capable of simultaneously explaining all of the parameters in the resulting data.

Like many others in cognitive science, van den Berg and colleagues assume that decisions are grounded in accumulating neural activity (which can be thought of as accumulating *evidence*, in this case evidence of direction of motion).[16] Accumulating activity in each of two neural populations (representing leftward motion and rightward motion respectively) takes a fluctuating "walk" toward a criterion set by the participants themselves in light of a speed / accuracy trade-off. It is assumed that evidence begins to accumulate in noisy fashion from a few milliseconds following stimulus onset until the decision-criterion is reached, with confidence-levels fixed by the extent of the difference between the two sets of signals (leftward versus rightward), together with reaction-time. Importantly, although van den Berg, Anandalingam, et al. (2016) arranged for the stimulus to disappear as soon as the joystick began moving following a decision, they assume that evidence continues to accumulate from information that has already entered the system but has not yet been processed. This continuing activity forms the basis for changes of mind about either the direction of motion or confidence. And it offers at least a partial explanation for why confidence and performance are less than perfectly correlated.

Note that the only representations that need to be appealed to in this account are first-order, world-directed, ones. Nothing meta-representational is required. In fact, it is the very same type of evidence that is used to drive initial decisions that also gets used to form the basis for high-risk or low-risk choices (confidence). On the one hand, there are graded (analog-magnitude) representations of leftward motion and rightward motion; the first of these to reach criterion determines the basic choice. But there is also (according to the authors' model) a representation of the *difference* in magnitude between these two representations, which (when integrated with reaction-times) is described by the authors as representing the odds, or probability, that the chosen direction of motion is, indeed, really the dominant motion. This is used to determine confidence (high-risk versus low-risk). And although the content of the confidence-representation *could* be stated in a way that makes it seem meta-representational (e.g. "the odds that the *representation* of leftward motion is correct"), in fact all it really represents is the probability that the overall motion in the stimulus is leftward (in a case where the larger analog magnitude represents leftward motion). Although this is computed from the difference in strength between two signals, it does not *represent* those signals—so it does not *meta*-represent. There is, as yet, no reason to postulate model-free metacognition here.[17]

The confidence-signal that issues in a choice of the high-risk or low-risk option does carry information about the relative rate of gain in two analog-magnitude representations, of course, representing leftward-motion and rightward-motion respectively. But we should again apply Shea's (2018) semantic framework to figure out the correctness condition for that signal. In order to do so, we need a better sense of the role that the signal plays in normal cognition. van den Berg, Zylberberg, Kiani, Shadlen, and Wolpert (2016) are able to demonstrate how it gets used in a *sequence* of decisions, for example, all of which must be correct for the intended outcome to be achieved and rewarded, but in the absence of feedback until after the final choice. Participants rely on the confidence-signal accompanying the first decision to alter the criterion-level (and hence their reaction time) for the second in a two-decision sequence. This is adaptive because it is *worth* investing extra time to get the second decision correct if you are confident of the first one, but not otherwise.

In light of this example, we can now ask what, from among the information carried by the confidence-signal, has stabilized its role in sequential choice (among other things, of course—this is just used for illustration). Plainly, the answer is that the strength of the confidence-signal correlates with the chance of making the first decision correctly. That it accurately carries information about the difference in strength between two competing signals (representing movement-left and movement-right) is of no use in itself. It is only because the magnitude of that difference correlates with the extent of the balance of evidence, and hence with the probability of having made a correct choice, that it is adaptive to modulate one's response to the second decision in the series. It is only adaptive to take more time over the second decision if the chances of having been correct in the first are high. So, the confidence-signal is best understood as a (subjective) representation of worldly probability or likelihood. When the choice made is for leftward-movement, the content of the confidence-signal says: high probability of leftward-movement; or perhaps: high probability of success when selecting the leftward-moving response.[18]

One notable feature of the account of implicit uncertainty behavior offered by van den Berg, Anandalingam, et al. (2016), is its claim that the very same drift-diffusion signals underlie both object-level responding (e.g. "leftward motion") *and* choice between the high-risk and low-risk options (that is, uncertainty-based behavior). The former signals are, of course, domain-specific ones. They are visual representations of patterns of movement. But Baer, Gill, and Odic (2018) and

[16] Such models are variously described in the literature as "sequential sampling with optional stopping," "drift diffusion," and "leaky competitive accumulator" accounts (Usher & McClelland, 2001; Pleskac & Busermeyer, 2010; Forstmann et al., 2016), and can take a variety of forms. Although there are some differences of detail among the various approaches, these need not concern us here.

[17] Barthelmé and Mamassian (2009), likewise, employ a behavioral task of a sort that could be used with animals (in which subjects select on every trial which of two stimuli they want to use to take the test), but they model subsequent performance using a Bayesian framework. They conclude that confidence can be thought of as the difference between two perceptual probabilities (probability of the leftward orientation of a Gabor patch minus probability of a rightward orientation). Again, probabilities are not meta-representational.

[18] Consistently with this suggestion that uncertainty responding is guided by estimations of worldly probability rather than anything metacognitive, Capuchin monkeys, who had previously been thought to be incapable of uncertainty-monitoring, turn out to make adaptive use of the opt-out option when the structure of the experiment is changed so that the odds of success when choosing randomly become one-in-six rather than one-in-two (Beran, Perdue, Church, & Smith, 2016). (Note that Capuchins famously have a try-everything-really-fast problem-solving style, whereas Macaques, that have more frequently been used as test subjects in binary-choice uncertainty-monitoring studies, are more cautious.)

Baer and Odic (2019) provide evidence (collected with children aged 3 to 7) that implicit certainty responding reflects a domain-general capacity, drawing on a common resource across different domains of judgment.

Specifically, Baer and Odic (2019) provide evidence that children's analog numerosity judgments develop independently of their capacity to respond adaptively to the certainty or uncertainty of such judgments. And Baer et al. (2018) demonstrate the same finding across the three domains of approximate number, approximate area, and degrees of emotionality in a face. Notably, in both studies children were presented with tasks that could equally well be used with non-human animals: they were presented side-by-side with *two* comparisons in a given domain and asked which one they wanted to choose for the test (to accumulate points if they were successful).

It is not obvious that these two perspectives are really inconsistent, however. Suppose it is true, as van den Berg and colleagues would suggest, that the same set of analog magnitude signals are employed *both* for object-level judgment (e.g. "*that* box has a greater number of dots") *and* for computing the certainty, or likelihood of correctness, attaching to that judgment. Still the computations involved are distinct. The object-level judgment depends on the accumulating but noisy neural signal (representing greater numerosity, say) together with the criterion for stopping. Certainty, in contrast, depends on computing the *difference* between two accumulating noisy signals (two signals of approximate number, in this case). It may be that this computation manifests a domain-general capacity—thereby enabling one to make cross-modal comparisons of likelihood—although the object-level judgments plainly do not.

The important point for our purposes, however, is that even if there is some tension between the two perspectives, from the fact that confidence-based decision making is a domain-general capacity, it doesn't begin to follow that it must involve meta-representational signals of any kind. On the contrary, it plausibly just involves the capacity to represent *likelihoods* (e.g. the likelihood that one approximate-number is larger than another, or the likelihood that one face is happier than another), mapping them to a common (domain-general) likelihood-scale where they can be compared and responded to.

6. Responding to ignorance

Claims to have discovered a form of model-based metacognition in animals have included not just the uncertainty-monitoring studies critiqued in Section 3, but also alleged findings of memory-monitoring, or *meta-memory* (Hampton, 2001, 2005), as well as *information-search* paradigms employed with primates (Krachun & Call, 2009; Rosati & Santos, 2016) and preverbal infants (Goupil, Romand-Monnier, & Kouider, 2016). Those claims, too, have been criticized by Carruthers (2017, 2018). But we will argue here that they nevertheless contain an important nugget of truth: underlying the capacity to utilize memory adaptively (and especially underlying an organism's responses to *failed* searches of memory), are forms of model-*free* metacognitive signal. And the same is true, too, of related capacities for epistemic emotions like curiosity and interest, as well as for instrumental information-search. We will begin by discussing curiosity, in particular.

Almost all philosophers and cognitive scientists who have written on the topic of curiosity have addressed it in metacognitive terms—as involving a desire *for knowledge* or true *belief*, or as an intrinsic motivation *to learn*, or something of the sort. (See Foley, 1987; Goldman, 1999; and Williamson, 2000, among philosophers; and see Litman, 2005; Gruber, Gelman, & Ranganath, 2014; Blanchard, Hayden, & Bromberg-Martin, 2015; and Kidd & Hayden, 2015, among psychologists. Exceptions among philosophers include Whitcomb, 2010, and Friedman, 2013, in addition to Carruthers, 2018.) Even Loewenstein's (1994) well-known "information gap" theory of curiosity, which sounds as if it might not require metacognition, is actually framed in metacognitive terms. Curiosity is said to arise from "a discrepancy between what one knows and what one *wishes to know*" (p.93; italics added).

Carruthers (2018) argues that this widespread view is mistaken. Rather, curiosity and interest are both species of *questioning attitude*. They are affective / motivating states that take questions rather than propositions as contents. Consider a monkey interested in (and closely observing) an on-going fight between two males in the troupe. On the standard view, the monkey has a metacognitive desire. The monkey *wants to know* who will win. Carruthers argues, in contrast, that the monkey has a desire-like state with a question as its content: *who will win?* On the standard account, the monkey must be supposed to possess the concept KNOW (or some rough equivalent). Carruthers' view is that this is unnecessary: the only concepts required for curiosity to be possible are concepts like WHO [will win], WHAT [that is], WHEN [food will come], and WHERE [food is], together with some concepts for kinds of action and kinds of thing (like winning and food).

Why should the questioning-attitude account be preferred to a metacognitive one? Primarily because of how widespread epistemic emotions are across the animal kingdom. Many animals (including rats and even bees) are motivated to explore novel or unrecognized environments (Cheeseman et al., 2014; Menzel et al., 2005; O'Keefe & Dostrovsky, 1971; Panksepp, 1998; Wills, Cacucci, Burgess, & O'Keefe, 2010). On the standard account they must be *wanting to know* what is around them. The contrasting view is that they are motivated by simple questions like, *what is around here?* Moreover, we know that the states at issue (in birds and mammals, at any rate) are at least curiosity-*like*, because individuals in both groups will (like humans) pay a proportion of any reward they might receive in a probabilistic reward-schedule just to know whether or not a reward is coming (Blanchard et al., 2015; Bromberg-Martin & Hikosaka, 2009; Fortes, Vasconcelos, & Machado, 2016; Gipson, Alessandri, Miller, & Zentall, 2009)—or in order to answer the question, *whether there will be food.*

Carruthers (2018) argues that a similar story can be told with respect to what motivates memory-search. Instead of treating it as motivated by a metacognitive desire (*wanting to remember*), we should understand it in terms of first-order questions directed at the animal's own memory system. (And to direct a query to one's memory system one doesn't need to have any conception of memory, of course; not even a simple one—the connection is surely built in, or hard-wired.) Consider a hungry food-caching bird, for example. Hunger prompts the question, *where food is*, which when directed toward the bird's own memory of cache locations either issues in an answer (*food is in that hole in that tree*)—in which case the bird flies there to retrieve it—or it does not (in which case the bird begins foraging-search afresh).

We assume going forward that these questioning-attitude accounts are correct. The contents of epistemic emotions like curiosity and interest are first-order questions directed at the environment, and the contents of memory searches are first-order questions directed at one's own memory. But what of the *causes* of such states? Epistemic emotions like curiosity and interest are caused by ignorance, or lack of knowledge. And animals behave differently depending on whether memory searches succeed or fail. Carruthers (2018) claims that being caused by ignorance is not the same as (meta-)representing ignorance, however. And that may well be true, if the kind of meta-representation at issue is the model-based variety. We now think, however, that the causes of such states may need to be explained in terms of model-*free* metacognition—that is, as involving concept-like signals that represent one's own ignorance in a model-free manner.

We should stress, however, that we are not suggesting that *all* forms of search behavior require (model-free) metacognition. Sometimes learning results from a random or semi-random walk through the environment. (The initial reconnaissance flights of bees when they first emerge from the hive may fall into this category.) And sometimes creatures might follow a gradient of information to a known target. For example, a male moth that detects the odor of a female faces a problem, since odor molecules are often distributed sparsely in turbulent air. Moths adopt a search strategy that involves flight across the up-wind

environment, computing the likely location of the female from occasional odor samples together with background assumptions about odor spread given the average wind direction and speed (Vergassola, Villermaux, & Shraiman, 2007; Voges, Chaffiol, Lucas, & Martinez, 2014). Nothing metacognitive is required here. In fact, each putative type of model-free metacognition will need to be examined on its own merits, and may require testing of alternative computational models of the data. Our goal here is just to motivate the question, and to make it seem plausible that at least some instances of model-free metacognition are real.

Consider, then, a cat that becomes curious when an unfamiliar mechanical toy is released and begins to move across the floor of the room.[19] The cat watches alertly, and may approach the object, perhaps sniffing and walking around it, or tapping it with a paw. According to the questioning-attitude view, the cat is motivated by the question, *what that is*. The likely account of how that question is prompted, or caused, proceeds as follows. When the object is first seen, neural populations representing a number of familiar entities become active, and start to compete with one another (in part through mutual inhibition). But at the same time neural activity builds, competitively, in a neural population that represents, *unknown*, or *unrecognized*. This reaches criterion first, issuing in the motivation to investigate—executing actions designed by evolution and/or previous learning to result in new knowledge. (These are actions that have often enough led to questions being *answered*—the question, in this case, being, *what that is*—thereby being rewarding, and reinforcing the actions in question.)

The suggestion, then, is that curiosity is always motivated by an explicit metacognitive signal. But in order for the signal in question to do its work, the creature need not possess even the most nascent model of the operations of its own mind; indeed, the same account will work just as well for bees, or for any creature that initiates question-answering behavior when in a position of ignorance. Hence our proposal, then, is that curiosity, interest, and related epistemic emotions are grounded in model-free forms of metacognition.

The account just sketched is based on the work of Dufau, Grainger, and Ziegler (2012), who offer a theory of the computations underlying human performance in a word / non-word task. They utilize a leaky competitive accumulator (LCA) model to explain how people swiftly categorize a new stimulus as a word (a familiar lexical item) or a non-word (an unknown or unrecognized sequence of letters). Neural activity representing *word* builds up over the course of milliseconds, depending on how closely the stimulus matches the features of a familiar word. If this activity reaches criterion one responds with "word." But at the same time the "non-word" response is linked to increasing activity in another population of neurons, from which activity in the "word" population subtracts. If this value reaches criterion level one answers "non-word."[20] A number of issues naturally arise from the attempt to deploy this sort of account to argue for model-free metacognition, however.

The first question to ask is why there needs to be an explicit signal representing, *unknown* (underlying curiosity) or, *non-word* (in the word-recognition task), leading one to respond accordingly. Perhaps the underlying mechanism in a word / non-word task can just be that evidence in favor of "word" builds more or less swiftly, and one answers "non-word" if that evidence doesn't reach the "word"-criterion *swiftly enough*. In effect, why can't the decision-criterion for "non-word" just be a temporal one? Perhaps the underlying mechanism can operate without any explicit signal representing, *non-word*. And then the mechanism underlying curiosity, likewise, can be that activity representing various familiar items builds in mutually-competitive fashion, with question-answering behavior initiated if *none* of them reaches criterion swiftly enough.

One answer to this challenge is that neural competition is ubiquitous in the brain, at all levels of organization (Mysore & Kothari, 2020). As a result, all forms of accumulator / drift diffusion / sequential sampling models of neural processing assume competition among options (Usher & McClelland, 2001; Pleskac & Busemeyer, 2010; Forstmann, Ratcliff, & Wagenmakers, 2016). Indeed, one of the main reasons why mutually-inhibitory competitive models have replaced ones that postulate an independent "race" to a decision criterion is that they enable an adaptive trade-off between speed and reliability (Teodorescu & Usher, 2013). This is because independent neural accumulation is much more influenced by noise (fluctuating neural activity), requiring a high decision-criterion to insure reliability, but at the cost of speed. So in the absence of a separate neural accumulator with the content, *unknown*, decisions to engage or not engage curiosity would likely be either unreliable or slow.

In addition, there is continual competition among motor plans as well (Cisek & Kalaska, 2010). Consider, then, the behavioral options available to the cat in our earlier example. These include stalking, leaping onto, biting, and so on, in addition to investigating; and these will be in competition with one another too, with the *investigate* option being implemented if it reaches criterion first. Note that the neural activity representing *investigate* also carries information about current ignorance, operationalized here to mean the failure of any of the object-type representations to reach criterion. In which case, even if it were the only explicit representation underlying questioning behavior, it should nevertheless be thought of as what Millikan (1995) calls a "pushmi-pullyu" representation, with a dual motoric / indicative content. And on the indicative side it seems that what it represents is: *unknown*.[21]

A second question to ask is why the neural signal underlying curiosity behavior needs to be assigned the content *unknown* (a meta-representational content), rather than an object-level conjunction of negative contents (*not a mouse, and not ball, and not a food-bowl, and …*). One answer appeals to explanatory generality. Consider the word / non-word task again. Here, too, one can ask why the signal that issues in a non-word response when it reaches criterion should be assigned the content, *not a word* rather than, *not "house," and not "louse," and not …*, for all the possible word-representations that have become activated and are competing with it. The answer is obvious: these would be different each time. Yet the explanation for why one responds "non-word" on all the different occasions when one does is surely the same: it is because one thinks (represents) that the stimulus is not a word. The same point holds in the case of curiosity. Many different sets of potential object-representations and event-representations will be active across different instances of curiosity; yet one should surely offer the same explanation in each case: the animal responds as it does because the item or event is unknown to it.

Moreover, and relatedly, it is only the metacognitive content that rationalizes and makes sense of the role of these signals in sustaining investigative behavior. Indeed, we can ask (in the spirit of Shea, 2018) what stabilized the role of the signals in evolution. The answer, plainly, is that it carries the information, *unknown*, not that it carries the

[19] Of course, not *every* new item or event—nor even every new item or event that violates a prior expectation of some sort—will evoke curiosity. There must first have been an appraisal of relevance to the animal's goals or values. Indeed, it seems that curiosity, like attention (Corbetta et al., 2008), depends on prior appraisals of the relevance of a stimulus. For further development of this point, as well as of some of the other material contained in this section, see Carruthers (2023).

[20] There is related single-neuron work showing that there are neurons in the prefrontal cortex of monkeys that fire when a stimulus is *absent*, as well as neurons that fire when it is present (Merten & Nieder, 2012).

[21] Consistent with these suggestions, Ahmadlou et al. (2021) find in a study of the neural bases of curiosity-driven behavior in mice that there is a small sub-population of neurons in media zona incerta, threshold-crossing activity of which drives deep investigative behavior. This region receives input about general arousal following simultaneous presentation of two objects, one familiar and one not, and from sensory regions.

information that the stimulus is not any of the things in a potentially open-ended conjunction of possibilities.

Thirdly, however, we can ask whether the signal that initiates search behavior in cases of curiosity or interest needs to have the content, *unknown* rather than the content, *novel* or *new*. Whether or not something is new to an organism is a function of previous encounters, and can be specified without mentioning anything psychological. And indeed, what *initiates* search behavior when confronted with novelty (often taking the form of an orienting response or shift of attention), can be a first-order, non-metacognitive, signal. But the same cannot be true of what *sustains* curiosity or interest thereafter. For by that point the object or event is no longer *new*, but can still be *unrecognized* or *unknown*.

Imagine one is strolling through a city woodland, not attending to the scenes around one. (One may be listening to a podcast through headphones.) Nevertheless, even in the absence of focused attention, one's visual system will produce a gist representation of the surroundings ("inanimate grass, bushes, and trees"). But now the visual system detects animate motion in a nearby tree. This issues in an error signal—similar to an "oddball" stimulus of the sort used in vision labs, but in this case with the content, *animate*. That will attract attention, and one is now consciously aware of the creature. But if the latter fails to be recognized, the result may be curiosity and sustained attention. It seems the content of the signal that sustains attention must be, *unknown* rather than, *new*, if only because the animal in question might have been seen multiple times before. ("What *is* that animal I keep seeing in these woods?")

Similar points can be made with respect to the signals underlying failed memory search. When the hungry bird directs a question at its memory with the content, *where a food cache is*, multiple neural populations representing potential locations will be evoked into initial activity, building in mutually-competitive manner to the extent that they contain relevant stored memory traces or are associatively activated by stored memory traces. But at the same time a neural population representing, *no memory* or, *unknown* will also begin building, competing with the specific-location representations to reach criterion first. If the signal representing, *no memory* wins the competition, then the bird sets off to forage afresh; if one of the location representations wins, on the other hand, then the bird sets off toward that spot.

There are the same general reasons, here—as there was in the case of failed recognition—to think that there must be an explicit neural signal representing memory failure, one that has *lack of memory* as its correctness condition. And the same general reasons also favor a metacognitive interpretation of the content of the signal, rather than a conjunction of negative claims (*not in tree A, and not in tree B, and not …*) which will vary from case to case. (Note that in this case there is simply no option corresponding to the proposed content, *new*, that might be thought to initiate curiosity.) Moreover, it is the absence of memory that has stabilized the role of the signal in question, adaptively initiating novel foraging behavior.

It is worth noting, however, that there need be nothing metacognitive in cases of *successful* memory retrieval. Suppose that the bird's memory-query results in a positive answer: *food is in tree A*. The bird doesn't need to represent this *as* a memory in order to act on it successfully. That the representation can now be used to guide a food-retrieval action is built into its functional role as a memory. And although the signal in question does carry the information that it derives from a memory system, it also carries the information that there is food stored at the location in question. Plainly, it is the latter type of information that has stabilized the role of memory in animal behavior. What makes the memory-signal correct is *not* that it derives from the creature's own memory (a meta-representational correctness-condition), but rather that there really is food at the location represented. This is a first-order, world-directed, content.

Moreover, even if we suppose that animals, too, can rely on the same sorts of cues that underlie human feelings-of-knowing to continue probing memory, nothing meta-representational need be involved here either. Specifically, partial retrieval, or retrieval of related items, will encourage people to continue a currently-unsuccessful memory search (Koriat, 1993; Koriat & Goldsmith, 1996; Koriat & Levy-Sadot, 2001). Suppose that animals rely on the same cues. But we already know that evaluative learning will lead creatures to value actions that take them closer to a goal. In this case the goal is to answer a question, *where food is*. If partial representations of food locations have led to full answers in the past when probing was continued, then the animal will be motivated to continue pressing the same question for a while. It doesn't have to feel or judge, *that is known*, in order for this to happen.

Our discussion so far in this section has assumed some form of competitive accumulator account of the processes underlying decision making. It might be wondered how the causes of curiosity might look from within a predictive-coding framework, however, of the sort that has been claimed by some to be ubiquitous at all levels of cognition (Clark, 2013, 2016; Friston, 2009, 2010). Although there have been some discussions of curiosity within this literature, in some cases the focus has been on responding, in the abstract, to curiosity as a possible counter-example to the thesis that the goal of all cognitive processes is error-minimization (Clark, 2018). And although some work modeling curiosity-like behavior has been done in this framework (Friston et al., 2017; Schwartenbeck et al., 2019), no one has, to the best of our knowledge, discussed curiosity following recognition failure, nor done the sort of detailed testing-and-modeling using reaction-time data of the kind conducted by Dufau et al. (2012).

On its face, however, it appears that a predictive-coding account of how curiosity gets caused in these circumstances would be the inverse of the competitive-accumulator one sketched earlier. As information about the unknown object is received, initial activity in a suite of predictive representations (MOUSE, BALL, FOOD-BOWL, and so on) will be *reduced* via bottom-up *error* signals. And then if the processes at this level are competitive ones, one might predict—from the same general consideration of trade-offs between speed and reliability—that competing activity in an alternative representation (UNKNOWN) would *not* be reduced by error signals, remaining as the only one standing. And it would be this that initiates (when deemed relevant enough) curiosity and exploratory behavior.

These considerations do not prove the case conclusively, of course. The issue is an empirical one. But error signals, like all other neural activity, will build over time in a noisy fluctuating manner. So mutually-inhibitory competition among such signals would serve to reduce wait-times for a decision while improving reliability, just as it does in accumulator models. And that means, in cases where the basic decision is between *unknown* and one or more active possibilities, that it would be beneficial to have a separate pairing of prediction and prediction-error representations of ignorance to compete with the remaining possibilities. In any case, however, the same basic form of experiment to be described in Section 8 could be employed as a test here too, provided the predictive models used are constructed around competitive, noisily fluctuating, levels of error-minimization.

In conclusion: since forms of curiosity, interest, and search behavior in general are extremely widespread in the animal kingdom; and since memory for locations, too, is equally widespread (in both cases extending to bees and other insects); we therefore have reason to think that the metacognitive signals whose existence we have proposed are model-free ones. One doesn't need even the simplest of theories of one's own mind in order to respond adaptively to failures of recognition and failures of memory-search. Yet the mechanisms underlying such responses will include explicit representational signals that contain facts about one's own mental states among their correctness conditions.

7. Mental effort

Carruthers (2021) argues that a model-free analog-magnitude signal representing the extent of executive engagement—or degrees of cognitive control—underlies decisions to engage, or not engage executive control (as well as affective feelings of mental effort in humans, at least).

That argument will be sketched here.

Controlled cognitive processing is generally experienced as aversive; and at its heart are capacities for top-down attentional control (Shipstead, Lindsey, Marshall, & Engle, 2014; Tsukahara, Harrison, Draheim, Martin, & Engle, 2020). Moreover, there is evidence, not only that controlled attention is present in many birds and mammals, at least (Karten, 2015; Mysore & Knudsen, 2013; Sauce, Wass, Smith, Kwan, & Matzel, 2014), but also that it is used in System-2-like forms of multistep planning (Gruber et al., 2019; Taylor, Elliffe, Hunt, & Gray, 2010; von Bayern, Danel, Auersperg, Mioduszewska, & Kacelnik, 2018). In which case many animals besides humans may take decisions to engage or disengage from controlled cognitive processing—but presumably represented in a model-free manner, as we argue shortly.

Decisions to engage control depend on a calculation of the *expected value of control* (Inzlicht, Shenhav, & Olivola, 2018; Shenhav et al., 2017; Shenhav, Cohen, & Botvinick, 2016). But it is possible that the decision itself need not involve any symbol-involving event that is a decision to engage cognitive control. That is to say, there need not be an explicit representation of cognitive control at the "decision"-point—not even a model-free one. Neural activity representing the different sources of information (costs, likelihood, and outcome value) may integrate and build in competition with alternatives, either reaching or failing to reach criterion. The result can be described as a decision to engage cognitive control (or not), but that result need not employ any symbolic structure or signal referring to control. When the criterion is met, control is engaged. But that this is *about* engaging cognitive control can be left implicit in the procedures involved.

Now, many animals are known to integrate the costs of *physical* effort into their decision making. They evaluate, not just the value of the end-state aimed at and the likelihood of achieving it, but also the energetic costs of getting there. And it is now known that rats, in addition to humans, will evaluate the costs of *mental* effort, too, with the evaluative networks involved being distinct from those that evaluate physical effort (Inzlicht et al., 2018; Winstanley & Floresco, 2016). The tasks that have been employed with rats involve a decision between two different task options, one requiring effortful focused attention to detect a briefly presented flash of light in one of five locations for a larger reward, the other only requiring the animal to detect an easily visible longer flash for a smaller reward. The animals have to trade-off the size of the reward against the attentional effort involved. Furthermore, rats can be conditioned to *positively* evaluate cognitive effort, at least within a particular domain or type of task (Hosking, Crocker, & Winstanley, 2016). So, too, can humans. Moreover, in humans, at least, evaluative conditioning can issue in a sort of learned cognitive industriousness that transfers to rather different types of cognitive-control task (Eisenberger, 1992). (Transfer of cognitive industriousness has not yet been tested in rats.)

The negative value attaching to controlled processing isn't fixed, then. Evaluative learning can change an animal's appraisal of the badness of expending cognitive effort in a given context. And then if a given form of cognitive effort is evaluated as bad (or good), it must first be represented. For affective systems can only appraise and evaluate items that are explicitly represented, as Delton and Sell (2014) point out. Indeed, all desires and emotions are *about* something, and result from prior affective valuation (or "appraisal") of the thing, event, or property in question.[22]

In many cases this means that concept-like representations are involved. If a monkey is to experience alarm at the sight of a snake, then it must have a concept-like representation of snakes. It must be capable of discriminating snakes from other things, for example, even if it knows very little about them. And affective learning, too, generally requires concept-like representations of the kinds in question. In order to acquire and store a positive or negative valuation of Xs, a creature must be capable of representing Xs in some fashion, and of distinguishing them from other types of thing.

What is surely true is that if controlled processing is to be negatively (and sometimes positively) evaluated, then it must be represented somehow. But it need not be represented conceptually, and certainly not in a model-based manner. Provided that the evaluative system receives a signal from executive controllers whenever the latter are engaged (and with the strength of the signal varying with the extent of that engagement), then it can be built into the default settings of the former that it should issue in negative affect. But normal processes of evaluative learning can change these settings, resulting in executive engagement in some contexts being evaluated positively. The signal in question refers to controlled processing, but without needing to be embedded in any theory-like representation of attention, cognitive control, or any other sort of mentality.

We can apply Shea's (2018) semantic framework to further establish the point. The signal received by evaluative systems when controlled cognitive processing is engaged carries the information that it has been so engaged. Moreover, that it carries that information explains how a dual-systems architecture has been stabilized by evolution. For given that attention is a limited resource, sustained attention to a thing or task carries opportunity costs, and should thus be negatively evaluated by default (Kurzban, 2016; Kurzban, Duckworth, Kable, & Myers, 2013). Hence it is generally adaptive to find controlled cognitive processing to be effortful. And it is *because* the strength of the signal in question co-varies with degrees of executive control that it plays the role that it does in computing the expected value of control, and in modulating ongoing cognitive processing.

Why should we think that the signal in question refers to cognitive control, however (something mental), rather than the underlying physical brain activity? Perhaps what is signaled is just increased activity in regions of prefrontal cortex, in particular. This suggestion might have made sense if the "ego depletion" model of mental effort had been correct. On this view, mental effort tracks calorific depletion in the brain resulting from frontal engagement (Masicampo & Baumeister, 2008). But this account is now thought to be problematic (Kurzban, 2010; Hagger et al., 2016; Vadillo, Gold, & Osman, 2016; Inzlicht et al., 2018). Instead, it is thought that feelings of mental effort result from the opportunity-cost of not directing attentional resources elsewhere (Kurzban, 2016). If this is correct, then the best explanation of how the role of effort-signals came to be stabilized in human and animal cognition needs to be pitched at the cognitive level; and in consequence, what they represent belongs at that level also.

It might be wondered, however, why explicit signals correlating with the extent of one's executive engagement need to be present at all. Why can't adaptive use of cognitive-control mechanisms depend on direct computations of opportunity costs instead? Perhaps the expected benefits of the current activity can be computed and compared with the expected benefits of what else one could be using one's attentional resources for at the time. However, these computations would themselves be cognitively costly. There is no end of alternative things one *could* be doing with the cognitive resources available, and even computations of the expected value to be derived from the most salient ones would be burdensome. This is, arguably, precisely why evolution has provided a sort of summary estimation of the opportunity costs, coded into a default value for mental effort—albeit one whose value can be adjusted by

[22] Perhaps some affective states can be "free-floating," with chemical causes rather than resulting from anything resembling a cognitive appraisal process. This might be true of moods like cheerfulness and depression, for example. (In fact, however, even moods are thought by some to reflect more-generalized appraisals of the opportunities afforded by the current environment; Eldar, Rutledge, Dolan, & Niv, 2016; Eldar, Pessiglione & van Dillen, 2021). But that cannot be what is going on here, since evaluative learning of mental effort is targeted specifically at executive control mechanisms.

learning and varies with the circumstances.[23]

It seems, then, that there are good reasons to think that cognitive effort depends on an analog-magnitude signal designating degrees of controlled processing. But why should we think that the signal in question is a model-free one, however? There are at least two reasons. The first is that it seems quite unlikely that all rats, mice, and birds should possess even a highly-simplified model of the operations of their own minds, or possess any concept-like representation of cognitive control as such. (After all, it remains controversial whether even monkeys are capable of model-based metacognition, as we noted in Section 2.) But the second reason is that no model-embedded representation of controlled processing *needs* to be present for the system to work as described. It can be built into the wiring and subsequent functioning of the affective system that the strength of the signal received as input from frontal systems represents the degree of engagement of controlled processing. Given the adaptive importance for an organism of making effective use of its limited cognitive-control abilities, it makes sense that this would evolve independently of, and prior to, any need to represent mental states as such. Moreover, given how widespread evaluations of cognitive effort are across mammalian (and probably avian) species, it is quite plausibly an evolutionary adaptation of just this sort.

8. Conclusion and path forward

The main goal of this paper has been to motivate a distinction between model-based and model-free kinds of metacognition. The former is the familiar target of all (or almost all) classic investigations of metacognitive capacities in humans, and is seemingly also the intended target of the search for metacognition in non-human animals. We have argued previously that metacognitive interpretations of most of the studies conducted with animals are unnecessary, and that the findings can be explained more simply in first-order terms (Carruthers, 2017; Carruthers & Williams, 2019; Ritchie & Carruthers, 2012). And we have recently presented positive evidence (summarized in Section 3) that at least some of the kinds of task employed with animals fail to engage model-based metacognitive capacities when used with humans (Nicholson et al., 2019, 2021). But that leaves open the possibility of model-*free* forms of metacognition in animals (as well as in humans, of course).[24]

Before making our positive case, we critiqued two possible candidates for model-free metacognition. In Section 4 we argued that the error signals involved in evaluative learning fail to have meta-representational contents. And in Section 5 we argued that the signals underlying implicit uncertainty-monitoring tasks represent object-level odds or likelihoods, rather than anything about the animal's own mind. But then in Section 6 we argued that forms of behavior that are an adaptive response to ignorance, on the one hand, and memory failure, on the other, are grounded in model-free concept-like signals with the content, *unknown*. And in Section 7 we argued that decisions to engage or not engage cognitive control are grounded in appraisals of an analog-magnitude signal representing the extent of that engagement. So we have proposed that there are at least two kinds of representation that have model-free metacognitive contents, representing ignorance and executive engagement respectively. (There may well be others, of course.)

How might these latter proposals be tested empirically? How one tests a theory depends on the relevant contrasting theories, of course. In the case of metacognitive signals of executive engagement, there are already well-established models of decision making that incorporate cognitive effort (Inzlicht et al., 2018; Shenhav et al., 2016; Shenhav et al., 2017). The relevant contrasting proposal, here, is that the signals are model-based rather than model-free. It has already been established that rats are capable of evaluating cognitive control (Hosking et al., 2016; Winstanley & Floresco, 2016), and one might think it implausible that rats should deploy any sort of model of the operations of their own minds, no matter how simplified. But the wider the extent of species that can be shown to be capable of evaluating cognitive effort, the more likely it will be that the signals in question are model-free ones. If this could be demonstrated for birds for example, or perhaps for vertebrates more broadly, then that would serve to make our hypothesis even more likely.

More-direct tests of model-based executive control in rats (say) would depend on the details of the proposed structure of the model. Any model-based view will predict additional flexibility in responding and intervening, of course. But this can't be tested directly until the proposed components and parameters of the model are specified. So there is a challenge, here, for anyone wishing to claim some form of self-awareness or model-based metacognition in rats or other creatures: specify the nature of the creature's proposed mind-model, and the place of executive control within it, and then predictions for flexible self-directed interventions can be derived and tested.

When discussing model-free signals of ignorance, in contrast, we considered two alternative theories. One is that curiosity depends just on the failure of any of the salient category-representations to reach criterion within a given time-frame, without there being any separate neural population that carries the information, *unknown*. The other is that there is such a neural population, but it represents, *new* rather than, *unknown*. In order to test the former, one might fit the two types of model against data from an object-classification task. Participants could be presented with a series of pictures of familiar and unfamiliar artifacts, classifying them into "kitchen implement," "office implement," or, "not known." Reaction times and error rates could be collected, and the data fitted against models that employ three noisy competing accumulators or just two. And one might control for the suggestion that the unrecognized items are represented as *new* rather than *not known* by insuring that all of those items had been seen during warm-up trials.

If modeling of this sort produced positive results, then that would support the existence of model-free metacognitive signals in humans and other animals. For of course many, many, creatures besides ourselves can respond adaptively to ignorance, and there is no reason to think that the mechanisms in humans should be any different in tasks of this sort. Admittedly, the discovery of model-free metacognition in nonhuman animals would likely fail to evoke the kinds of interest and excitement that have attached to claims of nascent (model-based) self-awareness in animals, or simple forms of self-consciousness in animals. Nor would its discovery in humans contribute to our understanding of the sort of self-awareness that we take to be so important about ourselves (Fleming, 2021). But these discoveries would, all the same, make valuable additions to our understanding of how human and animal minds function, and of the sorts of representations that underlie that functioning.

Funding statement

The research reported in Section 3 was supported in part by an *Economic and Social Research Council* (UK) Research Grant (ES/M009890/1) awarded to David M. Williams, Sophie E. Lind, and Peter Carruthers.

[23] This is likely to be the reason why stress has well-known negative effects on attentionally-demanding tasks (Liston, McEwen, & Casey, 2009). If one is under stress, then the expected benefit of paying more attention to one's surroundings (and hence of *not* focusing intently on the current task) should be ratcheted upwards.

[24] It also leaves open the possibility that there is model-based metacognition in animals that can be detected by other methods. Moreover, it is possible that the model-free metacognitive signals that exist in animals can readily *become* model-based when the animals in question acquire relevant forms of mentalizing ability. (This might be true of the signals of mental effort discussed in Section 7, for example, which now seem to be used by humans as cues when making explicit judgments about the accuracy of their own judgments, as we noted in Section 5.) For purposes of this paper we remain neutral on the question of which species of animal are capable of third-party mentalizing.

Author statement

Peter Carruthers: conceptualization; writing first draft; design of studies reported in Section 3.

David Williams: editing; design and execution of studies reported in Section 3.

Declaration of Competing Interest

None.

Acknowledgments

We are grateful to the following for their comments on earlier version of this article: Joe Gurrola, Chris Masciari, Shen Pan, Aida Roige, Samuel Warren, and a number of anonymous referees.

References

Ackerman, R., & Thompson, V. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21*, 607–617.
Ahmadlou, M., Houba, J., van Vierbergen, J., Giannouli, M., Giminez, G.-A., van Weeghel, C., … Heimel, J. A. (2021). A cell type-specific cortico-subcortical brain circuit for investigatory and novelty-seeking behavior. *Science, 372, 704*. eabe9681.
Baer, C., Gill, I., & Odic, D. (2018). A domain-general sense of confidence in children. *Open Mind: Discoveries in Cognitive Science, 2*, 86–96.
Baer, C., & Odic, D. (2019). Certainty in numerical judgments develops independently of the approximate number system. *Cognitive Development, 52*, 100817.
Bahrami, B., Olsen, K., Latham, P., Roepstorff, A., Rees, G., & Frith, C. (2010). Optimally interacting minds. *Science, 329*, 1081–1085.
Baillargeon, R., Scott, R., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology, 67*, 159–186.
Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology, 5*(9). e1000504.
von Bayern, A., Danel, S., Auersperg, A., Mioduszewska, B., & Kacelnik, A. (2018). Compound tool construction by new Caledonian crows. *Nature Scientific Reports, 8*, 15676.
Beck, J. (2018). Analog mental representation. *Wiley Interdisciplinary Reviews: Cognitive Science, 9*. e1479.
Beran, M., Perdue, B., Church, B., & Smith, J. D. (2016). Capuchin monkeys (*Cebus apella*) modulate their use of an uncertainty response depending on risk. *Journal of Experimental Psychology: Animal Learning and Cognition, 42*, 32–43.
Beran, M., Perdue, B., Futch, S., Smith, J. D., Evans, T., & Parrish, A. (2015). Go when you know: Chimpanzees' confidence movements reflect their responses in a computerized memory task. *Cognition, 142*, 236–246.
van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M., & Wolpert, D. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife, 5*. e12192.
van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M., & Wolpert, D. (2016). Confidence is the bridge between multi-stage decisions. *Current Biology, 26*, 3157–3168.
Bermúdez, J. (2003). *Thinking without words*. Oxford: Oxford University Press.
Blanchard, T., Hayden, B., & Bromberg-Martin, E. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron, 85*, 602–614.
Bromberg-Martin, E., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron, 63*, 119–126.
Brunsdon, M., & Happé, F. (2014). Exploring the "fractionation" of autism at the cognitive level. *Autism, 18*, 17–30.
Butlin, P. (2021). Cognitive models are distinguished by content, not format. *Philosophy of Science, 88*, 83–102.
Buttelmann, F., & Kovács, A. (2019). 14-month-olds anticipate others' actions based on their belief about an object's identity. *Infancy, 24*, 738–751.
Byrne, R., & Whiten, A. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.
Cammaerts, M. (2004). Operant conditioning in the ant *Myrmica sabuleti*. *Behavioral Processes, 67*, 417–425.
Camp, E. (2004). The generality constraint, nonsense, and categorical restrictions. *The Philosophical Quarterly, 54*, 209–231.
Carr, E., Rotteveel, M., & Winkielman, P. (2016). Easy moves: Perceptual fluency facilitates approach-related action. *Emotion, 16*, 540–552.
Carruthers, P. (2009a). Invertebrate concepts confront the generality constraint (and win). In R. Lurz (Ed.), *The philosophy of animal minds*. Cambridge: Cambridge University Press.
Carruthers, P. (2009b). 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: Oxford University Press.
Carruthers, P. (2014). Two concepts of metacognition. *Journal of Comparative Psychology, 128*, 138–139.
Carruthers, P. (2017). Are epistemic emotions metacognitive? *Philosophical Psychology, 30*, 58–78.
Carruthers, P. (2018). Basic questions. *Mind & Language, 22*, 130–147.
Carruthers, P. (2021). Explicit nonconceptual metacognition. *Philosophical Studies, 178*, 2337–2356.
Carruthers, P. (2023). The contents and causes of curiosity. *British Journal for the Philosophy of Science, 74*.
Carruthers, P., & Williams, D. M. (2019). Comparative metacognition. *Animal Behavior and Cognition, 6*, 278–288.
Casasanto, D., & Chrysikou, E. (2011). When left is "right": Motor fluency shapes abstract concepts. *Psychological Science, 22*, 419–422.
Cheeseman, J., Millar, C., Greggers, U., Lehmann, K., Pawley, M., Gallistel, C., Warman, G., & Menzel, R. (2014). Way-finding in displaced clock-shifted bees proves bees use a cognitive map. *Proceedings of the National Academy of Sciences, 111*, 8949–8954.
Cisek, P., & Kalaska, J. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience, 33*, 269–298.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–204.
Clark, A. (2016). *Surfing uncertainty*. Oxford: Oxford University Press.
Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences, 17*, 521–534.
Corbetta, M., Patel, G., & Shulman, G. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron, 58*, 3063–24.
Couchman, J., Coutinho, M., Beran, M., & Smith, J. D. (2009). Metacognition is prior. *Behavioral and Brain Sciences, 32*, 142.
Couchman, J., Coutinho, M., Beran, M., & Smith, J. D. (2010). Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition. *Journal of Comparative Psychology, 124*, 356–368.
Dayan, P., & Berridge, K. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 473–492.
Delton, A., & Sell, A. (2014). The co-evolution of concepts and motivation. *Current Directions in Psychological Science, 23*, 115–120.
Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning and Behavior, 22*, 1–18.
Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Stevens handbook of experimental psychology*. New York: John Wiley and Sons.
Dokic, J. (2012). Seeds of self-knowledge: Noetic feelings and metacognition. In M. Beran, J. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of metacognition*. Oxford: Oxford University Press.
Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four nonverbal measures. *Cognitive Development, 46*, 12–30.
Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief*. Oxford: Oxford University Press.
Dretske, F. (1988). *Explaining behavior*. Cambridge, MA: MIT Press.
Dufau, S., Grainger, J., & Ziegler, J. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1117–1128.
Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Washington DC: Sage Publications.
Eisenberger, R. (1992). Learned industriousness. *Psychological Review, 99*, 248–267.
Eldar, E., Pessiglione, M., & van Dillen, L. (2021). Positive affect as a computational mechanism. *Current Opinion in Behavioral Sciences, 39*, 52–57.
Eldar, E., Rutledge, R., Dolan, R., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences, 20*, 15–24.
Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.
Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906–911.
Fleming, S. M. (2021). *Know thyself: The science of self-awareness*. New York, NY: Basic Books.
Fleming, S. M., & Daw, N. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*, 91–114.
Fodor, J. (1990). *A theory of content and other essays*. Cambridge, MA: MIT Press.
Foley, R. (1987). *The theory of epistemic rationality*. Cambridge, MA: Harvard University Press.
Forgács, B., Parise, E., Csibra, G., Gergely, G., Jacquey, L., & Gervain, J. (2019). Fourteen-month-old infants track the language comprehension of communicative partners. *Developmental Science, 22*. e12751.
Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology, 67*, 641–666.
Fortes, I., Vasconcelos, M., & Machado, A. (2016). Testing the boundaries of "paradoxical" predictions: Pigeons do disregard bad news. *Journal of Experimental Psychology: Animal Learning and Cognition, 42*, 336–346.
Friedman, J. (2013). Question-directed attitudes. *Philosophical Perspectives, 27*, 145–174.
Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*, 293–301.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*, 127–138.
Friston, K., Lin, M., Frith, C., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Computation, 29*, 2633–2683.
Frith, U., & Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind & Language, 14*, 82–89.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*, 493–501.
Giere, R. (1988). *Explaining science*. Chicago: University of Chicago Press.
Gipson, C., Alessandri, J., Miller, H. C., & Zentall, T. (2009). Preference for 50% reinforcement over 75% reinforcement by pigeons. *Leaning & Behavior, 37*, 289–298.
Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*, 585–595.
Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy, 21*, 725–740.
Goldman, A. (1999). *Knowledge in a social world*. Oxford: Oxford University Press.
Goldman, A. (2006). *Simulating Minds*. New York: Oxford University Press.
Gopnik, A. (1993). The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16*, 1–14.
Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In P. Caruthers, S. Laurence, & S. Stich (Eds.), *The cognitive basis of science*. New York: Cambridge University Press.
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, 113*, 3492–3496.
Gross, J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry, 26*, 1–26.
Gruber, M., Gelman, B., & Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron, 84*, 486–496.
Gruber, R., Schiestl, M., Boeckle, M., Frohnwieser, A., Miller, R., Gray, R. D., Clayton, N., & Taylor, A. H. (2019). New Caledonian crows use mental representations to solve metatool problems. *Current Biology, 29*, 686–692.
Hagger, M., Chatzisarantis, N., Alberts, H., Anggono, C., Batailler, C., Birt, A., Brand, R., Brandt, M., Brewer, G., Bruyneel, S., Valvillo, D., Campbell, W., Cannon, P., Carlucci, M., Carruth, N., Cheung, T., Crowell, A., De Ridder, D., Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573.
Hampton, R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences, 98*, 5359–5362.
Hampton, R. (2005). Can Rhesus monkeys discriminate between remembering and forgetting? In H. Terrace, & J. Metcalfe (Eds.), *The missing link in cognition*. Oxford: Oxford University Press.
Hosking, J., Crocker, P., & Winstanley, C. (2016). Prefrontal cortical inactivations decrease willingness to expend cognitive effort on a rodent cost/benefit decision-making task. *Cerebral Cortex, 26*, 1529–1538.
Hyde, D., Simon, C., Ting, F., & Nokolaeva, J. (2018). Functional organization of the temporal-parietal junction for theory of mind in preverbal infants: A near-infrared spectroscopy study. *Journal of Neuroscience, 38*, 4264–4274.
Inzlicht, M., Shenhav, A., & Olivola, C. (2018). The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences, 22*, 337–349.
Jozefowiez, J., Staddon, J., & Cerutti, D. (2009a). Metacognition in animals: How do we know that they know? *Comparative Cognition and Behavior Reviews, 4*, 29–39.
Jozefowiez, J., Staddon, J., & Cerutti, D. (2009b). Reinforcement and metacognition. *Comparative Cognition and Behavior Reviews, 4*, 58–60.
Kammermeier, M., & Paulus, M. (2018). Do action-based tasks evidence false-belief understanding in young children? *Cognitive Development, 46*, 31–39.
Karten, H. (2015). Vertebrate brains and evolutionary connectomics: On the origins of the mammalian "neocortex". *Philosophical Transactions of the Royal Society B, 370*, 20150060.
Khalvati, K., Kiani, R., & Rao, R. (2021). Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nature Communications, 12*, 5704.
Kiani, R., Corthell, L., & Shadlen, M. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84*, 1329–1342.
Kidd, C., & Hayden, B. (2015). The psychology and neuroscience of curiosity. *Neuron, 88*, 449–460.
Király, I., Oláh, K., Csibra, G., & Kovács, Á. (2018). Retrospective attribution of false beliefs in 3-year-old children. *Proceedings of the National Academy of Sciences, 115*, 11477–11482.
Kobayashi, S., Kojima, S., Yamanaka, M., Sadamoto, H., Nakamura, H., Fujito, Y., Kawai, R., Sakakibara, M., & Ito, E. (1998). Operant conditioning and escape behavior in the pond snail, *Lymnaea stagnalis*. *Zoological Science, 15*, 683–690.
Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639.
Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.
Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 34–53.
Kornell, N., Rhodes, M., Castel, A., & Tauber, S. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*, 787–794.
Kornell, N., Son, L., & Terrace, H. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science, 18*, 64–71.
Krachun, C., & Call, J. (2009). Chimpanzees (*Pan troglodytes*) know what can be seen from where. *Animal Cognition, 12*, 317–331.
Kurzban, R. (2010). Does the brain consume additional glucose during self-control tasks? *Evolutionary Psychology, 8*, 244–259.
Kurzban, R. (2016). The sense of effort. *Current Opinion in Psychology, 7*, 67–70.
Kurzban, R., Duckworth, A., Kable, J., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences, 36*, 661–726.
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.
Liston, C., McEwen, B., & Casey, B. (2009). Psychosocial stress reversibly disrupts prefrontal processing and attentional control. *Proceedings of the National Academy of Sciences, 106*, 912–917.
Litman, J. (2005). Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition and Emotion, 19*, 793–814.
Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116*, 75–98.
Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, 1–17. https://doi.org/10.1093/nc/niw002
Masicampo, E., & Baumeister, R. (2008). Toward a physiology of dual-process reasoning and judgment. *Psychological Science, 19*, 255–260.
Menzel, R., Greggers, U., Smith, A., Berger, S., Brandt, R., Brunke, S., Bundrock, G., Hülse, S., Plümpe, T., Schaupp, F., Schüttler, E., Stach, S., Stindt, J., Stollhoff, N., & Watzl, S. (2005). Honey bees navigate according to a map-like spatial memory. *Proceedings of the National Academy of Sciences, 102*, 3040–3045.
Merten, K., & Nieder, A. (2012). Active encoding of decisions about stimulus absence in primate prefrontal cortex neurons. *Proceedings of the National Academy of Sciences, 109*, 6289–6294.
Millikan, R. (1989). Biosemantics. *The Journal of Philosophy, 86*, 281–297.
Millikan, R. (1995). ushmi-pullyu representations. *Philosophical perspectives, 9: AI. Connectionism and Philosophical Psychology*, 185–200.
Miyamoto, K., Trudel, N., Kamermans, K., Lim, M., Lazari, A., Verhagan, L., Wittmann, M., & Rushworth, M. (2021). Identification and disruption of a neural mechanism for accumulating prospective metacognitive information prior to decision-making. *Neuron, 109*, 1396–1408.
Mysore, S., & Knudsen, E. (2013). A shared inhibitory circuit for both exogenous and endogenous control of stimulus selection. *Nature Neuroscience, 16*, 473–478.
Mysore, S., & Kothari, N. (2020). Mechanisms of competitive selection: A canonical neural circuit framework. *eLife, 9*. e51473.
Nelson, T., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *26. The psychology of learning and information*. Cambridge, MA: Academic Press.
Nersessian, N. (2002). The cognitive basis of model-based reasoning in science. In P. Caruthers, S. Laurence, & S. Stich (Eds.), *The cognitive basis of science*. New York: Cambridge University Press.
Nichols, S., & Stich, S. (2003). *Mindreading*. Oxford: Oxford University Press.
Nicholson, T., Williams, D. M., Grainger, C., Lind, S., & Carruthers, P. (2019). Relationships between implicit and explicit uncertainty monitoring and mindreading: Evidence from autism spectrum disorder. *Consciousness and Cognition, 70*, 11–24.
Nicholson, T., Williams, D. M., Grainger, C., Lind, S., & Carruthers, P. (2021). Linking metacognition and mindreading: Evidence from autism and dual-task investigations. *Journal of Experimental Psychology: General, 150*, 206–220.
Odic, D., Lisboa, J., Eisinger, R., Olivera, M., Maiche, A., & Halberda, J. (2016). Approximate number and approximate time discrimination each correlate with school math abilities in young children. *Acta Psychologica, 163*, 17–26.
O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research, 34*, 171–175.
Panksepp, J. (1998). *Affective neuroscience*. Oxford: Oxford University Press.
Perner, J., & Ruffman, T. (1995). Episodic memory and autonoetic consciousness: Developmental evidence and a theory of childhood amnesia. *Journal of Experimental Child Psychology, 59*, 516–548.
Pleskac, T., & Busemeyer, J. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*, 864–901.
Proust, J. (2014). *The philosophy of metacognition*. Oxford: Oxford University Press.
Rhodes, M., & Castel, A. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*, 550–554.
Ritchie, J. B., & Carruthers, P. (2012). The emergence of metacognition: Affect and uncertainty in animals. In M. Beran, J. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of metacognition*. Oxford: Oxford University Press.
Rosati, A., & Santos, L. (2016). Spontaneous metacognition in Rhesus monkeys. *Psychological Science, 27*, 1181–1191.
Rupert, R. (2018). Representation and mental representation. *Philosophical Explorations, 21*, 204–225.
Sauce, B., Wass, C., Smith, A., Kwan, S., & Matzel, L. (2014). The external-internal loop of interference: Two types of attention and their influence on the learning abilities of mice. *Neurobiology of Learning and Memory, 116*, 181–192.
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.
Schwartenbeck, P., Passecker, J., Hauser, T., FitzGerald, T., Kronbichler, M., & Friston, K. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife, 8*. e41703.
Scott, R., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*, 237–249.
Seed, A., & Tomasello, M. (2010). Primate cognition. *Topics in Cognitive Science, 2*, 407–419.
Shea, N. (2014). Reward prediction error signals are meta-representational. *Noûs, 48*, 314–341.
Shea, N. (2018). *Representation in cognitive science*. Oxford: Oxford University Press.

Shenhav, A., Cohen, J. D., & Botvinick, M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience, 19*, 1286–1291.
Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J. D., & Botvinick, M. (2017). Toward and rational and mechanistic account of mental effort. *Annual Reviews in Neuroscience, 40*, 99–124.
Shipstead, Z., Lindsey, D., Marshall, R., & Engle, R. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language, 72*, 116–141.
Jones, C. R. G., Simonoff, E., Baird, G., Pickles, A., Marsden, A., Tregay, J., Happé, F., & Charman, T. (2018). The association between theory of mind, executive function, and symptoms of autism spectrum disorder. *Autism Research, 11*, 95–109.
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., Shields, W., & Washburn, D. (2003). The comparative psychology of uncertainty monitoring and meta-cognition. *Behavioral and Brain Sciences, 26*, 317–373.
Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition, 130*, 1–10.
Taylor, A., Elliffe, D., Hunt, G., & Gray, R. (2010). Complex cognition and behavioral innovation in new Caledonian crows. *Proceedings of the Royal Society B: Biological Sciences, 277*, 2637–2643.
Teodorescu, A., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review, 120*, 1–38.
Tsukahara, J., Harrison, T. L., Draheim, C., Martin, J. D., & Engle, R. (2020). Attention control: The missing link between sensory discrimination and intelligence. *Attention, Perception, and Psychophysics, 82*, 3445–3478.
Usher, M., & McClelland, J. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.
Vadillo, M., Gold, N., & Osman, M. (2016). The bitter truth about sugar and willpower: The limited evidential value of the glucose model of ego depletion. *Psychological Science, 27*, 1207–1214.
Vergassola, M., Villermaux, E., & Shraiman, B. (2007). "Infotaxis" as a strategy for searching without gradients. *Nature, 445*, 406–409.
Voges, N., Chaffiol, A., Lucas, P., & Martinez, D. (2014). Reactive searching and infotaxis in odor source localization. *PLoS Computational Biology, 10*(10). e1003861.
Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*, 655–684.
Whitcomb, D. (2010). Curiosity was framed. *Philosophy and Phenomenological Research, 81*, 664–687.
Whiten, A., & Byrne, R. (1997). *Machiavellian intelligence II: Extensions and evaluations*. Cambridge University Press.
Williams, D. M. (2010). Theory of own mind in autism: Evidence of a specific deficit in self-awareness? *Autism, 14*, 474–494.
Williams, D. M., & Happé, F. (2009). What did I say? Versus what did I think? Attributing false beliefs to self among children with and without autism. *Journal of Autism and Developmental Disorders, 39*, 865–873.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
Wills, T., Cacucci, F., Burgess, N., & O'Keefe, J. (2010). Development of the hippocampal cognitive map in pre-weanling rats. *Science, 328*, 1573–1576.
Winstanley, C., & Floresco, S. (2016). Deciphering decision making: Variation in animal models of effort- and uncertainty-based choice reveals distinct neural circuitries underlying core cognitive processes. *Journal of Neuroscience, 36*, 12069–12079.
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*, 283–307.