# *Explicit nonconceptual metacognition*

**Peter Carruthers**


**Philosophical Studies**
An International Journal for Philosophy
in the Analytic Tradition

ISSN 0031-8116

Philos Stud
DOI 10.1007/s11098-020-01557-1

Springer

Philos Stud
https://doi.org/10.1007/s11098-020-01557-1

# Explicit nonconceptual metacognition

**Peter Carruthers**[1]

Accepted: 16 September 2020


**Abstract** The goal of this paper is to explore forms of metacognition that have rarely been discussed in the extensive psychological and philosophical literatures on the topic. These would comprise explicit (as opposed to merely implicit or procedural) instances of meta-representation of some set of mental states or processes in oneself, but without those representations being embedded in anything remotely resembling a theory of mind, and independent of deployment of any sort of concept-like representation of the mental. Following a critique of some extant suggestions made by Nicholas Shea, the paper argues that appraisals of the value of cognitive effort involve the most plausible instances of this kind of metacognition.

**Keywords** Cognitive control · Effort · Error signal · Metacognition · Meta-representation · Nonconceptual

## 1 Introduction

Metacognition is generally defined in the field as "thinking about thinking" (Flavell 1979; Nelson and Narens 1990; Dunlosky and Metcalfe 2009). 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 commonly understood as restricted to thoughts about one's own thoughts and other mental processes. That is how it will be used here—at least initially. (The definition will be broadened shortly to include nonconceptual as well as thought-like representations of one's own mental states.)

✉ Peter Carruthers
pcarruth@umd.edu

[1] Department of Philosophy, University of Maryland, College Park, MD 20912, USA

Published online: 27 September 2020

For the most part psychologists have focused on the determinants and accuracy of a variety of explicitly-expressed metacognitive judgments (Dunlosky and Metcalfe 2009). These include judgments 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. But psychologists have also investigated the development of these capacities through childhood, as well as declines in those capacities in older adulthood. Philosophers have rarely engaged with this literature directly. (Proust 2014, is one exception.) Instead, they have developed more-general theories of self-awareness and self-knowledge, debating to what extent it is, or is not, distinctively different in kind from our knowledge of the mental states of other people, for example (Bilgrami 2006; Carruthers 2011; Fernández 2013; Cassam 2014; Byrne 2018).

There have also been extensive debates about the phylogenetic origins of self-awareness. Some have claimed that the so-called *uncertainty-monitoring* and *memory-monitoring* tasks employed with monkeys manifest at least simple forms of metacognitive awareness of their own mental states, either as such, or in a way that preadapts the representations in question to become components in full-blown self-awareness in humans (Smith et al. 2003, 2014). Some critics have charged that these findings can be explained 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 (surprise, curiosity, uncertainty, feelings of knowing, and so on) are likewise first-order (non-metacognitive) in nature, quite different from the explicit judgments in humans that those feelings can ground (Carruthers 2017; Nicholson et al. 2019). Yet others have described the epistemic feelings in question as *procedurally* metacognitive because of the role they play in modulating and controlling ongoing cognitive processes, while denying that meta-representation of any sort is involved (Dokic 2012; Proust 2014).

This paper will introduce and evaluate a different possibility, exemplified in (and hitherto only in) the work of Shea (2014). This is that there might be metacognitive mental phenomena that are explicitly meta-representational in character, but which neither represent mental states *as such*, nor in ways that display any form of nascent self-awareness. These would be representational states whose correctness conditions involve the existence and/or properties of some mental state or process in the agent, but without being embedded in even a simple kind of understanding of the mental status of the states referred to. They would be explicit nonconceptual meta-representational states, referring to mental states in oneself.

I should stress that by *explicit* meta-representation, here and in what follows, I do not mean meta-representations that are conscious. (The explicit/implicit contrast is sometimes intended to line up with the conscious/unconscious one; but not here.) Rather, I mean that it involves some form of meta-representational symbolic structure or signal, as opposed to being built tacitly into the processing rules or procedures employed. To illustrate, the information that seeing leads to knowing is *ex*plicitly represented by the mentalizing system 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 *im*plicitly 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."

It should also be emphasized that the question of explicit *nonconceptual* metacognition requires that the definition of "metacognition" be weakened. It will need to encompass not just *thoughts* about one's own mental states and processes (which are presumably conceptual in nature) but potentially any kind of mental symbol that has some of one's own mental states among its correctness conditions, no matter how "low level" the functional role of that symbol might otherwise be. On this broadened understanding, provided that it is a representation of some sort, and that it represents some mental state or process in oneself, then it can qualify as metacognitive. Notice, however, that there are still many forms of self-monitoring that are excluded by this broadened definition, including the representations that participate in temperature regulation and balance regulation, and that monitor glucose levels in the bloodstream. For these are representations of *bodily* states and processes, not mental ones.

Section 2 will make some remarks about the idea of nonconceptual representation in general, and nonconceptual metacognition in particular. Sections 3 and 4 will then evaluate, and critique, Shea's (2014) case for believing in the meta-representational nature of the reward-prediction error signals that underlie evaluative learning, as well as the motor-control prediction errors that modulate action. (Note that both kinds of prediction error are extremely widespread in the animal kingdom, and are found even in insects. No one would claim that they constitute nascent forms of self-awareness.) Then Sects. 5 and 6 will discuss the burgeoning recent literature on goal-directed thinking and mental effort in humans and other animals, arguing that affective evaluations of so-called "controlled processing" depend upon nonconceptual meta-representational signals.

## 2 Nonconceptual representation

Nonconceptual representations, as I here understand them, are those that are fine-grained and continuous (or "analog"), as opposed to marking a categorical boundary of some sort (or "chunked"). This way of drawing the distinction between conceptual and nonconceptual representation is pretty standard in the philosophical literature (Tye 2000; Bermúdez 2015; Beck 2019), and has been familiar at least since Peacocke (1992). Thus, thinking that ripe tomatoes are red is a purely conceptual representation, composed of the concepts RIPE, TOMATO, and RED. In contrast, perceiving a roundish-shaped object whose surface is covered with some specific shades of red (but without conceptualizing the object *as* a red tomato), is a purely nonconceptual representation.

Note that representation *as* has traditionally been thought to require concepts (Fodor 2015), in which case representing a mental state *as such* would require deployment of some mental-state concept. Burge (2010) has argued for an important exception to this principle, however. He thinks that there can be nonconceptual representations of objects and objectivity *as such* in perception, in virtue of those perceptual states being designed to track perceptual constancies. Because perception

is designed to zero in on the same object irrespective of differences in orientation and lighting, he argues that it represents that thing *as* an object in an objective world. However, as we will see, there is nothing resembling a perceptual constancy in the putative cases of nonconceptual metacognition that will concern us. Hence these will (if genuine) involve representations of some of one's own mental states, but without representing them *as mental*.

There is a natural objection to the suggestion that nonconceptual representation (at least in the cases that interest us) can be understood as the obverse of representation *as* (where representing *as* requires deployment of concepts). This is that animals surely represent items in the world in ways that aren't purely analog in nature. Yet many philosophers deny concepts to animals, insisting that concepts must fully satisfy the "generality constraint" on concept-recombination (Evans 1982; Bermúdez 2003; Camp 2004). Rather than dispute the latter point, I prefer to formulate the idea of representation *as* in terms of concept-*like* entities. An animal can deploy a concept-like representation when it is capable of distinguishing instances of different kinds from one another, and when the concept-like state serves as a node for collecting information about instances of the kind. An animal might distinguish flying predators from climbing predators, and distinguish both from ground-based ones, for example; while storing separate bodies of information about instances of the three kinds tied to the three representations in question (Seyfarth et al. 1980). So representation *as*, as I understand it, at most requires concept-*like* states that can be components of belief-*like* ones. We can then remain agnostic about whether animals have genuine concepts and beliefs, while nevertheless claiming that they can represent things as such.

The sorts of (putative) nonconceptual meta-representation that will concern us are all instances of analog magnitude representation. The latter have been extensively investigated in the domain of analog number representation, in both humans and other animals (Dehaene 1997; Jordan et al. 2008; Izard et al. 2009). But there are also distinct forms of analog magnitude representation for area, density, length, and time (Odic 2018). What is distinctive of all analog magnitude representations, however, is that they obey Weber's law (Beck 2015). This states that just-noticeable differences in a magnitude maintain a constant ratio as the extent of the magnitude increases. Thus, an infant who can discriminate between magnitudes of five items and ten items but not between five items and eight items, will be able to notice the difference between ten items and twenty items, but not between ten items and sixteen items.

Our question, then, is whether there is anything like this in the metacognitive domain. Are there any explicit analog representations that have some of one's own mental states or one's own mental processes among their correctness conditions, but where these metacognitive states are not only nonconceptual in nature themselves, but also function and perform their role without deployment of any concept-like representation of the mental?

## 3 Reward-prediction error signals

Shea (2014) argues at length that reward-prediction error signals have these properties. 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.

Shea (2014) seems to suggest that a meta-representational view is implicit in the science of affective learning. Indeed, one might think that even the language used by theorists in the field—*error* signal—suggests as much, implying that what is signaled is that a representation (the content of a prediction) is erroneous or mistaken. Here (roughly) is how evaluative learning works[1]: an expected value for a given item or action is compared with the actual value experienced when that item is consumed or the action is performed. When the two don't match (the experienced value is higher or lower than expected) a reward-prediction error signal is generated, causing the stored value for things or actions of that kind to be updated by an increment dictated by a learning-rule. Shea asserts that the content/correctness-condition for the error signal *d* is that "the reward received for the last action was higher/lower than the currently-represented expectation for that action" (p. 322). Since the correctness-conditions include reference to both *expectation* and *representation*, the error signal qualifies as meta-representational, Shea claims.[2]

It is not obvious that the nomenclature used—*error* signal—is anything more than a theorist's external gloss, however. Of course, we as theorists can see 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*]."

[1] The simplified model presented here ignores top-down influences on evaluative learning, of the sort that figure in nocebo and placebo effects. It also ignores the influence of background mood (Eldar et al. 2016), and glosses over the fact that a value acquired from previous evaluative learning will be a *weighted* average of the values experienced in the past, with more recent rewards being counted more heavily in proportion to the learning rule.

[2] Notice that Shea does *not* claim that the error signal is meta-representational on the grounds that it represents something about an *action* (that is, a movement caused by an intention, hence a partly mental entity). This is for good reason: the action-representations involved in evaluative learning are coded entirely in sensorimotor format.

On the latter reading, an error signal represents the difference between two *values*, not two representations of value.

To elaborate on this alternative reading, here is a neutral (albeit still simplified) description of what happens when an animal is engaged in evaluative learning. The creature has coded in affective memory a stored value-magnitude ($m$) for items or actions of a certain kind $X$, which we can represent as: VAL($Xs = m$).[3] (This might either be an innate default setting of the evaluative system, or previously learned through affective conditioning.) Here "VAL" stands for "value" or "goodness/badness." But it is intended only as a theorist's tool, signifying something that is implicit in the overall operations of the system in question (namely, that $m$ is a magnitude *of value*); "VAL" is not itself a representation in the mind of the organism. (Compare the way in which people commonly use "BEL" to represent the attitude of believing something. If one says that an organism has the attitude BEL[$p$], the only representation attributed to the animal is the proposition $p$.) Thus, the only representations actually involved are representations of a kind of thing or action and a representation of a magnitude (of value).

Now, when the creature is about to encounter or enact a new instance of the kind, the stored value gives rise to an expectation of value with the content: VAL(*this* $X = m$); that is to say, it assumes that the value of the up-coming $X$ will be the same as the value of $Xs$ generally. The creature then consumes the item or engages in the action, and experiences a new value $n$, which is (let us suppose in this instance) greater than $m$. We can represent this as: VAL(*this* $X = n$), where $n$ equals $m + d$. The difference between $m$ and $n$ (that is, $d$, the error signal) is then used to update the stored value of $Xs$ in general in proportion to some learning rule $a$. So the new stored value of $Xs$ in general becomes: VAL($Xs = m + d/a$).

The simplest way to think of the content of the error signal is that it is an analog magnitude representation with the content *plus-d* (or in cases where the experienced value of an $X$ is less than the stored value, *minus-d*). What it represents is a difference in value between a specific instance of kind $X$ and the value of $Xs$ in general (up to that time or as previously encountered, derived either from the evolutionary history of the organism, or from previous evaluative learning, or both). It is implicit in the functioning of the entire system that the *plus* or *minus* components of the error signal represent the fact that the value of the current item is greater, or lesser, than the value predicted on the basis of previous experience. And the represented magnitude is then used to adjust the stored value for items of that kind accordingly, as dictated by a learning rule. The only representations participating in the entire learning process are a representation of a kind of item or action, a representation of a specific item or action of that kind, representations of

[3] Thus far Shea (2014) doesn't disagree. For on p. 322 he writes that the correctness-condition for a stored value (in my notation, $m$) is that, "[$m$] is accurate iff the average reward payoff that would be achieved by repeatedly choosing [the thing or action] in the current environment is [$m$]." Note the implication here, by the way, that the stored value for Xs is at least implicitly tensed. The system is designed to track values in potentially changing circumstances. In fact, it can be helpful to think of VAL($Xs = m$) as a sort of *generic* representation, the evaluative equivalent of a generic belief like, "Birds fly." Although learned from previous experience, it gives rise to expectations of future Xs, somewhat as the generic belief that birds fly would lead you to expect that the next bird you encounter will fly.

value attaching to the kind and to the specific instance, and a representation of the magnitude (and valence) of the difference between the two values. Nowhere do *expectations* or *experiences* need to figure among the contents represented. Nothing here needs to be meta-representational in nature.

Moreover, we can make perfectly good sense of the overall design of the affective-learning system without needing to introduce anything meta-representational into the explanation. All animals live in a changing world. The intrinsic (organism-relevant) values attaching to things often change over time. Fruit ripens (becoming better to eat) and then decays (becoming worse). And the probabilities of acquiring things of value can change with time, too. The likelihood of finding ripe fruit on a particular kind of tree will vary with the seasons, for example. In the simplest form of evaluative learning (known as "model free"; Dickinson and Balleine 2002; Dayan and Berridge 2014) these dimensions (outcome value and likelihood) are not tracked independently of one another. Rather, the stored value attaching to an action-in-a-context is incrementally shifted up or down (or left untouched) with each experience of reward or failure to experience reward. Since the world is not just change-ridden but also noisy, it would make no sense (it would not be adaptive) for the stored value of an action-type to be fully altered with each experience. (Although a particular sampled item of fruit may be rotten, most of the remaining items might be ripe. And although a particular tree may as yet be devoid of fruit, others might not be.) Nor would it make sense for the system to ignore an unexpected reward or failure of reward, even though it might, indeed, be just noise. The simplest, most efficient, way to track changing values in a changing but noisy world is to make incremental shifts with each experience. Hence learning rules have evolved, shifting the stored value up or down incrementally with each experience.

Another way to approach the same overall conclusion is to apply elements of the "varitel" semantic approach developed by Shea (2018) himself.[4] This arguably provides us with the most plausible account of representation in cognitive science. According to Shea, in many cases what fixes the specific content of a representation from among the disparate bodies of information that it carries, is the information that played a causal role in stabilizing the operations of the system in question, and ultimately the behavior of the organism. In general, he thinks causal stabilization can be done by evolution, by learning, or by contributing to individual survival. In the present case (the information carried by reward-prediction error signals), we presumably need to look to evolution, since organisms don't learn how to do evaluative learning, nor do they acquire a capacity for it during their lifetimes.

[4] Note that in his 2018 book Shea still explicitly endorses a meta-representational account of reward-prediction error signals (albeit in passing), so it isn't anachronistic to employ his later theory of representational content to evaluate the earlier account. Note, too, that Shea's varitel semantics 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. It is the first of these sorts or 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.

Rather, evaluative learning is among the basic learning mechanisms of almost all living creatures.

Now, the error-signal *plus-d* carries the information that an expectation of value had previously been formed while a greater value was subsequently experienced. So it does carry metacognitive information. But it also carries information about the difference in value between instances of a kind of thing that may have been encountered previously in 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. 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 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, without metacognitive content.

Shea (2014) attempts to develop a similar what-matters-in-evolution argument to the one just employed, but draws the opposite (metacognitive) conclusion. He writes (p. 332):

> We are supposing that the system has been set up the way that it is, by evolution or learning, in order to maximize overall average payoffs to the agent. It is the correlational information carried by *d* about [the] difference between represented expectation and feedback that contributes to achieving this overall outcome. So that correlation explains why *d* is wired up to be processed in the way that it is.[…] that is evidence that *d* is representing that the reward was more/less than the represented expected value and telling downstream processing to revise expected values accordingly […]. That content partly concerns the content of another of the system's representations [the expected reward], and so is meta-representational.

But it is *not* the correlation between *d* and a difference between two mental representations that explains why *d* is wired up to have the effects that it does. Rather, what has stabilized the system is the correlation between *d* and the difference in adaptive value between a kind of thing or action in general and the value of the current item.[5]

Does the account of evaluative learning sketched here depend upon a representational theory of value, however, of the sort defended by Cutter and Tye

[5] Note that in the quoted passage Shea describes the error-signal *d* as *telling* the down-stream system to revise its expected value for the thing in question. He intends this quite seriously. He thinks that the error signal has imperative, or *directive*, content as well as indicative content. (That is, he thinks it is what Millikan 1995, calls a "pushmi-pullyu" representation.) But this is ill-motivated. The error signal no more has an imperative content than does visual perception of something unexpected. The perceptual content serves to update one's beliefs about the likelihood of events in the environment. But it doesn't *direct one* to update one's beliefs.

(2011) or Carruthers (2018)? Does it presuppose that stored reward-values and experienced rewards are representations of the adaptive value of the associated things and actions? A critic might suggest, instead, that reward-values are intrinsic properties of the evaluative system that become associated with (and are evoked by) types of thing and types of action, and that play a certain functional role in the life of the organism (motivating pursuit and avoidance behavior, in particular). These intrinsic properties would have to be magnitudes of some sort, and they would have to allow for both positive and negative valences. But now the error signal, in representing the difference between two such magnitudes, would be representing a difference between two intrinsic mental properties. Hence the error signal would not be meta-*representational* (since the reward-values in question aren't representations, on this view), but it *would* be meta-*mental* (since reward-values are mental properties).

Shea's (2018) own semantic framework can again be deployed to show that reward-states are, indeed, representations of adaptive value, however. To see this, it will be helpful to recall the familiar distinction between so-called *primary* and *secondary* reinforcers (and punishers). Primary reinforcers are things, properties, or actions that have been fixed by evolution to motivate the actions that issue in them. (Primary punishers are things, properties, or actions that have been fixed by evolution to inhibit actions that cause them.) They include such things as eating when hungry, drinking when thirsty, and pain. Secondary reinforcers and punishers, in contrast, are things, properties, or actions that have acquired positive or negative value through their association with, or capacity to predict, the occurrence of a primary reinforcer or punisher. Thus, subsequent to evaluative learning that the sound of a bell predicts food, an animal may work to make that bell sound for its own sake, in the absence of any other reward. Likewise, following training an animal might work to silence a bell that had previously been predictive of an electric shock.

Mental states caused by primary reinforcers (positive rewards) carry information about properties that contribute directly to survival, reproduction, or both. And the causal role of positive rewards (motivating actions to obtain those reinforcers) has been stabilized through evolution precisely because of the adaptive value of the reinforcers themselves (food, drink, and so on). So they qualify as representations of adaptive value. (Likewise, punishment states like pain are representations of adaptive disvalue.) Mental states caused by secondary reinforcers and punishers (which are also states of reward and punishment) carry information that is predictive of primary reinforcers and punishers, and hence are likewise representations of adaptive value and disvalue. Since this is so, our earlier argument stands: the content of a reward-prediction error signal is just the magnitude of the difference between two values. The representation is first-order, not metacognitive.[6]

[6] It is worth noting that reward-value representations aren't *just* representations of adaptive value and disvalue; they are actually what Millikan (1995) calls "pushmi-pullyu" representations. For expectations of value directly motivate actions designed to achieve or avoid the valued or disvalued things in question. Moreover, note the difference between this case and the error signals themselves, which Shea (2014)

## 4 Motor-prediction error signals

Shea's (2014) claim that reward-prediction error signals are instances of nonconceptual metacognition is unconvincing, then. But before concluding this aspect of our discussion we should consider the plausibility of another claim he makes in that paper (albeit in passing). This is that *motor*-prediction error signals are also meta-representational in nature. For he might be right about this, even if wrong about the error-signals resulting from reward prediction.

It is now well established that whenever a creature generates a motor instruction to produce a planned movement, an efference copy of that instruction is used to create a "forward model" of the sensory feedback that should be received if the movement proceeds as planned (Wolpert and Kawato 1998; Wolpert and Ghahramani 2000; Grush 2004; Jeannerod 2006). The forward model is compared with the organism's proprioceptive and visual experience as the action unfolds, issuing in error-signals in cases of mismatch. This leads the organism to adjust the motor instructions accordingly. This sort of control architecture is quite ancient, and is employed even in dragonflies (Mischiati et al. 2015). So if the error-signals in question could be shown to have meta-representational contents, then this would surely qualify as an instance of explicit nonconceptual metacognition. Shea's view is that the error-signal represents the difference between a *predicted* (planned) movement and the *experienced* movement, and so carries meta-representational content about the (in)correctness of the organism's expectations. But there is little reason to accept this view.

Consider the simplest possible case. A young child has learned how to use her arm to push wooden blocks away from her along a flat surface. (Perhaps she is playing a game with a care-giver in which they exchange blocks with one another at a constant rate.) In light of that experience, she now generates a motor instruction similar to those used previously to push away an additional block. Suppose that the force generated by this instruction is $f$. The instruction is also used to create a forward model of the way things will look and feel as the block moves away from her with the usual velocity $v$. But in fact, the block is heavier and/or more friction-prone than previously, and the block only moves with velocity $w$. When matched against the content of the forward model, this issues in an error-signal with the content $v - w = x$, where $x$ is then the difference in velocity between the two velocity-magnitudes. As a result, the motor instructions are ramped upwards to generate a force $f + g$, where $g$ is the estimated additional force needed to result in velocity $v$. Here the error-signal represents the difference between two velocities: one that was expected and one that was experienced. But that they were expected and experienced doesn't need to be (meta-)represented. For this is implicit in the causal workings of the system itself. So there is no reason, here, to postulate explicit meta-representation of mental states of oneself.

Footnote 6 continued
claims have imperative content. For it is not true that *anything* that causes a change in an organism (e.g. in a stored value) is an imperative. Imperatives serve to cause / motivate *action*.

I conclude the discussion thus far, then, with the claim that reward-prediction and motor-prediction error signals are not promising places to look for instances of explicit nonconceptual metacognition. Shea (2014) deserves credit for having been the first to introduce the latter idea into the literature, however. For there may well be other regions of cognitive functioning where metacognitive signals of this sort can genuinely be found. We begin to explore this suggestion next, approaching it initially through the dual-systems literature. The suggestions made in Sect. 5, although seemingly plausible, will fail to pan out. But they provide the foundation for a related claim that will get vindicated in Sect. 6.

## 5 Deciding to think

A dual-systems framework has dominated psychological theorizing about the architecture of human cognition for well over two decades (Evans and Over 1996; Sloman 1996; Metcalfe and Mischel 1999; Stanovich 1999; Kahneman 2015). On this view, the human mind is comprised of two systems, or types of system. System 1 comprises a set of systems that are fast, parallel, automatic, and effortless. These issue swiftly in intuitions about correct answers when presented with reasoning problems, as well as in fast emotional responses and "gut feelings" to situations generally. System 2, in contrast, is slow, serial, controlled, and effortful. This is the system employed when we stop to "think through" a solution to a problem, or attempt to moderate our own emotional response to something.

The two systems are generally thought to be in competition with one another, with occasional switches happening from System 1 intuitive mode to System 2 controlled processing or vice versa. Indeed, the dominant model of the relations between the two systems is that cognition operates in System 1 mode by default, but with System 2 remaining active in the background. System 2 monitors the outputs of System 1, ready to intervene, inhibit those responses, and initiate controlled processing when needed (Evans 2010; Evans and Stanovich 2013). Having completed (or failed at) the task, people will then generally lapse back into effort-free System 1 intuitive processing once again.

From a common-sense perspective what one does, in many cases of switching from System 1 to System 2, is *decide* to stop and think, either about the answer to a question or about the correct solution to a problem. People will explain how they tackled a particular issue (in System 2 mode) by saying something like, "I thought that it looked a bit tricky, so I decided to stop and think it through before answering." Some people are habitually thoughtful, of course, and might engage System 2 processing as a matter of course, without any decision-making process. (Psychologists refer to this personality trait as "need for cognition"—see Cacioppo and Petty 1982.) But in many cases one might detect that one is confronted with a problem-type that one has failed at in the past, or one might intuit that the question

being asked is somehow difficult or not straightforward. Here something resembling a decision to switch to System 2 processing seemingly needs to be appealed to.[7]

Now, notice that a decision to stop and think about something is a decision with a meta-representational content. What one is deciding to do is engage a specific sort of executively-controlled cognitive processing. So the decision is *about* (and is designed to cause) mental states or processes in oneself, whether represented as such or nonconceptually.

Similar considerations might lead one to postulate a metacognitive decision in the opposite direction, when people switch from task-focused processing to mind-wandering (Christoff et al. 2016). For the competition between System 2 and System 1 processing can be understood (at least in part) as a competition between focused, controlled, uses of top-down attentional systems, on the one hand, and the so-called *saliency* system, on the other (Corbetta and Shulman 2002; Corbetta et al. 2008). The latter continually monitors unattended perceptual and mnemonic contents for relevance to current goals and values, causing a switch in the focus of top-down attention when those contents are deemed important enough or relevant enough. Some have argued that these switches of attentional focus result from *decisions* to redirect attention (Carruthers 2015). Although the things that are deemed to be relevant or irrelevant are the worldly contents of unattended percepts or memories, still a decision to redirect attentional resources would seem to have at least a nonconceptual metacognitive content. For the content of what one decides is to switch *that* [the top-down attentional system] to a novel topic.

One can question the relevance for our discussion of these decisions to engage or disengage from System 2 thinking. For they seem to be too "high level" to qualify as instances of the kind of explicit nonconceptual metacognition that forms our target of inquiry. Indeed, they are embedded in fully-conceptual forms of self-awareness, as well as capacities to represent controlled thinking *as such*. The issue is worth pursuing, however, because System 2 thinking is but one instance of controlled cognitive processing, at the heart of which are capacities for top-down attentional control (Shipstead et al. 2014; Tsukahara et al. 2020). And there is evidence, not only that controlled attention is present in many birds and mammals, at least (Mysore and Knudsen 2013; Sauce et al. 2014; Karten 2015), but that it is also used in System-2-like forms of multistep planning (Taylor et al. 2010; von Bayern et al. 2018; Gruber et al. 2019). So many animals, too, may take decisions to engage or disengage from controlled cognitive processing—but presumably represented nonconceptually, somehow.

In both of the above types of case, however (deciding to engage or disengage from controlled processing) it is possible to doubt whether any explicit decision-like event is really needed. In fact, there need be no event in such cases that involves a symbol or symbolic structure of any sort with metacognitive content. These can,

[7] Notice that in cases where one consciously and reflectively does something decision-like—such as articulating in inner speech, "I will stop and think about this one"—a decision has *already* been taken to engage controlled processing. (In pausing to articulate those words one is *already* stopping to think.) The inner-speech performance serves as an expression of that decision. My target in the discussion that follows are the (putative) *unconscious* decisions that initiate controlled processing.

rather, be instances of merely procedural metacognition (Proust 2014). This is because competitive processes of the sort appealed to here can potentially be explained through the use of leaky competitive accumulator (LCA) models of decision making (Usher and McClelland 2001), as I will now explain.[8]

LCAs are now widely employed across cognitive science. For example, they have been used to account for perceptually-based decision making, such as swiftly deciding whether a presented stimulus is a word or a non-word (Dufau et al. 2012). Neural activity representing WORD builds up over the course of milliseconds, depending on how closely the stimulus matches the features of a familiar word, modulated by ever-present spontaneous fluctuations in neural activity (neural noise; Boly et al. 2007; Hesselmann et al. 2008). If this activity reaches criterion quickly enough one responds with "word." But at the same time the "non-word" response is linked to a fixed value of activity in another population of neurons, from which the activity among the "word" population subtracts, via inhibition. If this value *doesn't* drop below criterion quickly enough one answers "non-word."

According to such LCA models, the criterion-level for a given decision is fixed by task demands and/or features of background motivation. For instance, if the experimenter in a word/non-word task emphasizes accuracy, then the criterion-level for a "word" response will be set high (with the criterion for "non-word" set correspondingly low), with consequent effects on one's reaction times. If the experimenter stresses speed of responding, in contrast, then those criteria can be shifted accordingly. But they will also be influenced by factors specific to the individual (either personality-like traits or results of previous learning). The important point for our purposes, however, is that the event that takes place when the neural activity that represents WORD hits criterion level, resulting in initiation of the action of saying "word", is not itself a symbol of any sort. While the "word" and "non-word" responses are explicitly represented (as are properties of the stimulus), the decision between them is implicit in the operations of a leaky competitive accumulator system.

If these models are applied to the "decision" to stop and think or the "decision" to divert attention from a task and begin mind-wandering, then we can offer non-metacognitive explanations of the phenomena. We just have to suppose (as is widely assumed already in cognitive science) that there is continual and active competition between the two systems (System 1 versus System 2, or top-down attention versus the saliency system), with criterion levels for "deciding" between them set by contextual goals, individual personality, and previous learning.

A similar framework can then be applied to theoretical models of dual-system decision-making that make use of the idea of the *expected value of control* (Shenhav

[8] To be clear, I will not be claiming that LCA models actually succeed in providing the *best* explanation of the phenomenon we call "deciding to think / stop thinking." My claim, rather, is negative. It is that, given the viability of such models and their popularity in psychology, it would be hard to establish—and certainly controversial to claim—that there is an explicit metacognitive representation-type picked out by the phrase "decision to engage controlled processing." Note, too, that although LCA models are a specific type of diffusion decision model (Forstmann et al. 2016), nothing of significance turns on this distinction for our purposes.

et al. 2016, 2017; Inzlicht et al. 2018). On this sort of account, people (and other animals) decide how much mental effort/controlled processing to invest in a given task by computing the expected value of doing so. That is, from previous experience of tasks of that sort they weigh the expected benefits (the likelihood of success combined with the value of a successful outcome) against the costs. People generally experience controlled processing to be aversive, probably resulting from the opportunity costs of not engaging those resources elsewhere, or of not allowing attention to spread more broadly (Kurzban et al. 2013). In any given case, an appraisal of that cost is weighed against the expected benefits to issue in a decision to engage control (and by how much), or not.

Here too, as in our previous discussion, there might not be 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 nonconceptual one. Neural activity representing the different sources of information (costs, likelihood, and outcome value) is integrated and builds, either reaching or failing to reach criterion within a given time-frame. 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.

I conclude that although common-sense might appeal to metacognitive decisions when explaining why someone stops to think about a question before answering, and although some in philosophy have employed similar language when explaining what causes someone to shift into mind-wandering mode, such appeals aren't mandatory. And they may well fail to qualify as the best explanations of the phenomena. Rather, any "decisions" involved in such cases can be left implicit in the outcome of competition among leaky competitive accumulators of various sorts. Nevertheless, as we will see next, embedded within such competitive processes are ones that really do require reference to cognitive control—specifically, those involved in evaluating the cost of control.

## 6 Mental effort

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. 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 (Winstanley and Floresco 2016; Inzlicht et al. 2018). 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

task-type or domain (Hosking et al. 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.)

In light of these findings it seems highly likely that evaluations of cognitive effort and controlled processing are widespread across mammalian species, at any rate. But given the detailed neural-wiring homologies between mammals and birds (Karten 2015), and given that some birds are known to be capable of controlled cognitive processing when solving problems (Taylor et al. 2010; von Bayern et al. 2018; Gruber et al. 2019), it may well be the case that birds, too, will assign value (more generally disvalue) to cognitive effort. I will now argue that such evaluations must involve (at least) a nonconceptual form of explicit meta-representation.

As is familiar, controlled processing (maintaining focused attention, response inhibition, and so on) is generally experienced as aversive. Thinking and focusing can be hard work. But the negative value attaching to controlled processing isn't fixed, as we have just seen. 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 be represented. For affective systems can only appraise and evaluate items that are explicitly represented, as Delton and Sell (2014) argue.

All desires and emotions are about something, and result from prior affective valuation (or "appraisal") of the thing, event, or property in question.[9] 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.

I suggest that all affective evaluation of a thing, property, or event requires representations of that thing, property, or event. Indeed, this was implicit in our discussion of reward-prediction errors in Sect. 3. The evaluative system stores a value-magnitude associated with a type of thing or action, creates expectations for the reward to be received from a given instance of that type, and subsequently creates an evaluative experience tied to that thing or action when it is consumed or performed, comparing it to the expected value. One can't assign a value to Xs without representing Xs in some fashion. As we will see, however, in order for controlled processing, in particular, to be evaluated, it need not be represented *as such*, nor even as a form of mentality. It can be represented as a nonconceptual *that*,

[9] Note, however, that this isn't to claim that all affective states in general are tied to a representation of something. Moods, in particular, are affective states that are free-floating—or, perhaps better, that color *everything*—rather than being tied to some thing or type of thing in particular.

with degrees of magnitude of *thatness* tracking degrees of engagement of controlled processing.

What is surely true is that if controlled processing/System 2 thinking is to be negatively (or sometimes positively) evaluated, then it must be represented somehow. But it need not be represented as such. Provided that the evaluative systems receive 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 they should issue in negative affect. But normal processes of evaluative learning can change these settings, resulting in some forms of executive engagement in some contexts being evaluated positively. The signal in question refers to controlled processing, but without needing to be embedded in any concept-like representation of attention, cognitive control, or any other sort of mentality.[10]

We can now apply Shea's (2018) varitel semantics 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 such information explains how the 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 et al. 2013). Hence it is generally adaptive to find controlled cognitive processing to be effortful. And it is *because* 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 to the underlying 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 (Masicampo and Baumeister 2008). But this view has now been thoroughly de-bunked (Kurzban 2010; Hagger et al. 2016; Vadillo et al. 2016; Inzlicht et al. 2018). Instead, it has been suggested that mental effort signals the opportunity-cost of not directing attentional resources elsewhere (Kurzban et al. 2013; 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.

[10] Note that if different forms of executive engagement are to be evaluated separately, as I hint at here (e.g. focused attention versus response inhibition), then the model I am proposing would require there to be distinct signals sent to evaluative systems from each component kind of executive control. The default settings for each of these signals would be negative, but evaluative learning might alter their values in particular types of context independently of one another. Note, too, that there is unlikely to be anything resembling perceptual constancies in this domain. (Mental effort doesn't have to be identified across a wide range of differing signals; a single signal, or a single signal for each type of effort, will do.) So one cannot appeal to Burge's (2010) framework to argue that despite the absence of mental-state concepts mental effort is represented *as such*.

What reasons are there for thinking that the explicit signals designating degrees of controlled processing are *nonconceptual* ones, however? There are at least two. The first is that it seems quite unlikely that rats, mice, or 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 explicit metacognition, as we noted in Sect. 1.) And indeed, hardly anyone in the field of comparative metacognition has claimed that rats are capable of thinking about their own mental processes. (However, see Kirk et al. 2014; and Templer et al. 2017.) But the second reason is that no concept-like 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 systems that when they receive that signal as input, it represents the 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.

We can conclude, then, that Shea (2014) is partly correct. The best place to look for explicit nonconceptual metacognition is, indeed, in the evaluative domain. But the most plausible example isn't the reward-prediction error signal, as he claims. Rather, it is in the evaluation of controlled processing undertaken by many creatures besides ourselves. This requires evaluative systems to receive an analog-magnitude signal referring to such processing (and whose strength co-varies with the extent of that processing), but without such processing needing to be represented *as such*, and without any concept-like representations of mental states or processes needing to be deployed.

## 7 Conclusion

This paper has explored the question whether there are forms of explicit (not merely procedural) reference to an agent's own mental states or processes that occur independently of any concept-like representation of those states or processes. I have argued (against Shea 2014) that reward-prediction and motor-prediction error signals do not qualify for this sort of explicit but nonconceptual metacognitive status. But I have also argued that affective evaluation of controlled processing of various kinds likely *does* qualify. In humans the negative evaluation inherent in feelings of cognitive effort may generally be experienced as effortful *thinking* or effortful *attending*. That is to say, humans can experience the effort of cognitive control as such. But this isn't needed for the dual-systems framework to function as intended. In fact, in both humans and other many other creatures, engagement of cognitive control may issue in an analog-magnitude signal transmitted to affective systems. The meaning of the signal is fixed by evolution and implicit in the operations of the overall network. It refers to cognitive control (and to the extent of cognitive control) without that control being represented as such. But the upshot

(negative valence caused by controlled processing) is correctly classified in humans, at least, as effortful thinking or attending, and can thus be represented as such. I suggest that the analog magnitude signal itself, however, should be seen as an instance of explicit nonconceptual metacognition.

**Acknowledgments** I am grateful to the anonymous referees for their insightful comments on previous versions of this article.

## References

Beck, J. (2015). Analogue magnitude representations: A philosophical introduction. *British Journal for the Philosophy of Science, 66*, 829–855.

Beck, J. (2019). Perception is analog: The argument from Weber's law. *Journal of Philosophy, 116*, 314–349.

Bermúdez, J. (2003). *Thinking without words*. Oxford: Oxford University Press.

Bermúdez, J. (2015). Nonconceptual mental content. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/fall2015/entries/content-nonconceptual

Bilgrami, A. (2006). *Self-knowledge and resentment*. Cambridge, MA: Harvard University Press.

Boly, M., Balteau, E., Schnakers, C., Degueldre, C., Moonen, G., Luxen, A., et al. (2007). Baseline brain activity fluctuations predict somatosensory perception in humans. *Proceedings of the National Academy of Sciences, 104*, 12187–12192.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Byrne, A. (2018). *Transparency and self-knowledge*. Oxford: Oxford University Press.

Cacioppo, J., & Petty, R. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116–131.

Camp, E. (2004). The generality constraint, nonsense, and categorical restrictions. *Philosophical Quarterly, 54*, 209–231.

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

Carruthers, P. (2015). *The centered mind*. Oxford: Oxford University Press.

Carruthers, P. (2017). Are epistemic emotions metacognitive? *Philosophical Psychology, 30*, 58–78.

Carruthers, P. (2018). Valence and value. *Philosophy and Phenomenological Research, 97*, 658–680.

Cassam, Q. (2014). *Self-knowledge for humans*. Oxford: Oxford University Press.

Christoff, K., Irving, Z., Fox, K., Spring, N., & Andrews-Hanna, J. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature Reviews Neuroscience, 17*, 718–731.

Corbetta, M., & Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*, 201–215.

Corbetta, M., Patel, G., & Shulman, G. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron, 58*, 3063–3124.

Cutter, B., & Tye, M. (2011). Tracking representationalism and the painfulness of pain. *Philosophical Issues, 21*, 90–109.

Dayan, P., & Berridge, K. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, and Behavioral Neuroscience, 14*, 473–492.

Dehaene, S. (1997). *The number sense*. London: Penguin Press.

Delton, A., & Sell, A. (2014). The co-evolution of concepts and motivation. *Current Directions in Psychological Science, 23*, 115–120.

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: Wiley.

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.

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*. Thousand Oaks: Sage Publications.

Eisenberger, R. (1992). Learned industriousness. *Psychological Review, 99*, 248–267.

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.

Evans, J., & Over, D. (1996). *Rationality and reasoning*. Hove, East Sussex: Psychology Press.

Evans, J. (2010). *Thinking twice: Two minds in one brain*. Oxford: Oxford University Press.

Evans, J., & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*, 223–241.

Fernández, J. (2013). *Transparent minds: A study of self-knowledge*. Oxford: Oxford University Press.

Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906–911.

Fodor, J. (2015). Burge on perception. In S. Laurence & E. Margolis (Eds.), *The conceptual mind*. Cambridge: MIT Press.

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.

Gruber, R., Schiestl, M., Boeckle, M., Frohnwieser, A., Miller, R., Gray, R. D., et al. (2019). New Caledonian crows use mental representations to solve metatool problems. *Current Biology, 29*, 686–692.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27*, 377–442.

Hagger, M., Chatzisarantis, N., Alberts, H., Anggono, C., Batailler, C., Birt, A., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573.

Hesselmann, G., Kell, C., & Kleinschmidt, A. (2008). Ongoing activity fluctuations in hMT+ bias the perception of coherent visual motion. *Journal of Neuroscience, 28*, 14481–14485.

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.

Inzlicht, M., Shenhav, A., & Olivola, C. (2018). The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences, 22*, 337–349.

Izard, V., Sann, C., Spelke, E., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences, 106*, 10382–10385.

Jeannerod, M. (2006). *Motor cognition*. Oxford: Oxford University Press.

Jordan, K., MacLean, E., & Brannon, E. (2008). Monkeys match and tally quantities across senses. *Cognition, 108*, 617–625.

Kahneman, D. (2015). *Thinking, fast and slow*. New York: Farrah, Strauss, & Giroux.

Karten, H. (2015). Vertebrate brains and evolutionary connectomics: On the origins of the mammalian "neocortex". *Philosophical Transactions of the Royal Society B, 370*, 20150060.

Kirk, C., McMillan, N., & Roberts, W. (2014). Rats respond for information: Metacognition in a rodent? *Journal of Experimental Psychology: Animal Learning and Cognition, 40*, 249–259.

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.

Masicampo, E., & Baumeister, R. (2008). Toward a physiology of dual-process reasoning and judgment. *Psychological Science, 19*, 255–260.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: The dynamics of willpower. *Psychological Review, 106*, 3–19.

Millikan, R. (1995). Pushmi-pullyu representations. *Philosophical Perspectives: AI, Connectionism and Philosophical Psychology, 9*, 185–200.

Mischiati, M., Lin, H.-T., Herold, P., Imler, E., Olberg, R., & Leonardo, A. (2015). Internal models direct dragonfly interception steering. *Nature, 517*, 333–338.

Mysore, S., & Knudsen, E. (2013). A shared inhibitory circuit for both exogenous and endogenous control of stimulus selection. *Nature Neuroscience, 16*, 473–478.

Nelson, T., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and information*. Cambridge: Academic Press.

Nicholson, T., Williams, D., 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.

Odic, D. (2018). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science, 21*, e12533.

Peacocke, C. (1992). *A study of concepts*. Cambridge, MA: MIT Press.

Proust, J. (2014). *The philosophy of metacognition*. Oxford: Oxford University Press.

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.

Seyfarth, R., Cheney, D., & Marler, P. (1980). Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behavior, 28*, 1070–1094.

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., et al. (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.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.

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.

Stanovich, K. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah: Erlbaum.

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.

Templer, V., Lee, K., & Preston, A. (2017). Rats know when they remember: Transfer of metacognitive responding across odor-based delayed match-to-sample tests. *Animal Cognition, 20*, 891–906.

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*,

Tye, M. (2000). *Consciousness, color, and content*. Cambridge: MIT Press.

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.

von Bayern, A., Danel, S., Auersperg, A., Mioduszewska, B., & Kacelnik, A. (2018). Compound tool construction by New Caledonian crows. *Nature Scientific Reports, 8*, 15676.

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.

Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience, 3*, 1212–1217.

Wolpert, D., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*, 1317–1329.