PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Metacognition and reasoning

Logan Fletcher and Peter Carruthers

Phil. Trans. R. Soc. B 2012 **367**, 1366-1378
doi: 10.1098/rstb.2011.0413

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 1366–1378
doi:10.1098/rstb.2011.0413

Research

Metacognition and reasoning

Logan Fletcher and Peter Carruthers*

Department of Philosophy, University of Maryland, College Park, MD 20742, USA

This article considers the cognitive architecture of human meta-reasoning: that is, metacognition concerning one's own reasoning and decision-making. The view we defend is that meta-reasoning is a cobbled-together skill comprising diverse self-management strategies acquired through individual and cultural learning. These approximate the monitoring-and-control functions of a postulated adaptive *system* for metacognition by recruiting mechanisms that were designed for quite other purposes.

Keywords: control; metacognition; meta-reasoning; monitoring; reasoning; system 2

1. INTRODUCTION

Mature humans normally possess some degree of ability to manage their own first-order mental activities. They can monitor and control the processes involved in learning and remembering, for example, as well as some of the thought processes involved in reasoning and decision-making. Thus, we sometimes pause to consider whether we are reasoning about some matter in an appropriate way, or to wonder whether we might be unduly influenced by some extraneous factor in arriving at a decision. An important question concerns the cognitive underpinnings of these metacognitive abilities. When we reflect on our own processes of learning or reasoning, are we employing a distinctive cognitive system specifically adapted for just this kind of reflective self-management? Or are we rather (as we will argue) using a hodge-podge of self-management techniques picked up from our culture and our individual experience, enlisting whatever cognitive mechanisms are available to be pressed into service?

Metacognition of learning and remembering has been extensively studied by psychologists [1–9]. The consensus in the field is that metacognitive monitoring is inferential rather than direct, and is grounded in a variety of sensorily accessible cues. For example, a feeling of knowing can be grounded in the familiarity of the cue or recall of facts closely related to the target item, and a judgement of learning can be based on the fluency with which the item to be learned is processed. While these findings might be consistent with the claim that there is nevertheless an evolved system or 'module' for metacognition, in fact, there is no reason to believe that any cognitive mechanism is employed other than the same mindreading faculty that we use for attributing mental states to other people, enhanced with some acquired first-person strategies and recognition abilities [10]. Indeed, when brain scans of people engaged in such metacognitive activities are conducted (and appropriate first-order tasks are used for purposes of subtraction), the very same network of regions that has been found to be involved in mindreading tasks is seen to be active. This includes medial prefrontal cortex, posterior cingulate cortex and the temporo-parietal junction [11–13]. Moreover, although a number of studies have found brain areas outside the mindreading network to be active during metacognitive tasks, these have employed metacognitive *accuracy* as their main comparison measure [14], and this will, of course, involve more than mere mindreading. The accuracy of one's metacognitive judgement that one has just answered a question correctly, for example, will depend upon a great many factors in addition to one's capacity to attribute thoughts to oneself [10].

Despite this apparent consensus among metacognition researchers, many in surrounding disciplines within cognitive science continue to be committed to the idea of a specific adaptation for metacognition with direct (as opposed to sensory-cue-based) access to its domain. Thus, a number of comparative psychologists have claimed to find evidence of metacognitive abilities in primates who seem to lack equivalent mindreading abilities [15–23]. They suggest that the evolution of these first-person monitoring-and-control abilities would have provided the foundation for the later evolution of third-person mindreading [24]. Likewise, a number of accounts of human self-knowledge capacities require the support of just such an evolutionary theory, which is needed to explain the emergence of their postulated mechanisms of direct access to our own mental states [25–27]. Moreover, many of those in computer science and artificial intelligence who build metacognitive systems assume that these systems should incorporate direct access to the processes being monitored, as well as mechanisms for direct intervention and control [28]. And similar assumptions are made by some of those who explain conscious forms of cognition in terms of metacognition, who go on to claim that the evolutionary function of this kind of consciousness is metacognitive control [29].

* Author for correspondence (pcarruth@umd.edu).

One contribution of 13 to a Theme Issue 'Metacognition: computation, neurobiology and function'.

Although the literature on human metacognition mentioned earlier is in some tension with—or at least fails to support—such views, this literature has focused almost exclusively on metacognition of learning and remembering, rather than metacognition of reasoning and decision-making. (Indeed, the standard textbook in the field of metacognition does not discuss the latter at all; see [9].) So it might be claimed that the evidence of an adaptive mechanism for metacognition may be found in this domain instead. It is for this reason that we focus on reasoning in our discussion. We will show that metacognition of reasoning, too, is a cobbled-together skill lacking the expected signature properties of an adaptation. Throughout, we will understand reasoning broadly, to encompass not only processes that take us from factual premises to a factual conclusion, but also processes that take us from some set of desires and beliefs to a decision or action. This is essentially the same as the understanding adopted by Kahneman *et al.* [30] in the studies that launched the extensive literature on heuristics and biases in human reasoning.

Psychologists who study reasoning have increasingly come to accept some or other version of 'dual-systems' or 'dual-process' theory, which posits two different systems involved in human reasoning, each with its own characteristic cluster of features [31–39]. These are generally now referred to as 'system 1' and 'system 2'. While not everyone accepts the legitimacy of a two-systems account of reasoning and decision-making, at least in any strong form [40,41], what is at stake for our purposes is not the claim that system 1 processes are always either associative or heuristic, while system 2 processes are always rational (as some have claimed). On the contrary, on the account that we propose, system 2 reasoning depends in part on one's beliefs about norms of reasoning, and so will only be as rational as one's beliefs are (which might well involve consulting the entrails of a chicken rather than performing a Bayesian calculation). As we will see, what matters for our purposes is just the distinction between forms of reasoning that depend importantly on mental rehearsal and the resources of working memory and those that don't. And on this, there seems to be no substantive disagreement [42].

System 1 is now widely believed to encompass a *set* of different systems that operate in parallel, delivering swift and intuitive judgements and decisions in response to perceptual inputs in a manner that is unconscious, automatic and guided by principles that are, to a significant extent, innately fixed and universal among humans. System 2 reasoning, on the other hand, is conscious and reflective in character, and is thought to proceed in a slow, serial manner, according to principles that vary among both individuals and cultures.

Three additional features of system 2 are particularly relevant to the topic of meta-reasoning. First, system 2 is generally held to be subject to intentional control. Second, it can be guided by normative beliefs about proper reasoning methods. And third, one of the principal roles often attributed to system 2 is to override the unreflective responses that are issued automatically by system 1 in reasoning tasks, when these fall short of appropriate standards of rationality. As we will see, such system 2 interventions will sometimes proceed by initiating episodes of controlled, reflective, reasoning that serve to displace and supersede the original system 1 reasoning. In other cases (especially in affectively based decision-making), system 2 resources are deployed in order to *modulate* or *suppress* system 1 processing, by intentionally reframing or directing attention away from the sensory representations that comprise the inputs to system 1.

Taken together, these features would appear to give the dual-systems set-up just the right structural profile to serve as a locus for the sort of monitoring-and-control functions postulated by those who believe in a distinct adaptation for metacognition. In order for system 1 to be overridden, it might seem that it needs to be subject to some sort of metacognitive monitoring to assess the rational adequacy of its intuitive verdicts. The 'control' component of the meta-reasoning system would then be realized through the normatively guided use of system 2 to displace, modulate or suppress system 1. On the hypothesis of a distinctive adaptation for meta-reasoning, then, one should predict that people will reliably display a robust native competence for overriding system 1 when it fails to satisfy normative standards, as well as for acting to suppress or redirect it, and for switching over to reflective system 2 processing, as appropriate. In addition, since system 2 is under intentional control, and is also widely thought to be imbued with metacognitive awareness (given that it operates consciously), one would expect the presence of a meta-reasoning system to be manifested in a reliable natural facility in noticing and correcting errors in one's system 2 reasoning procedures as well. For these are just the functions that the hypothesized meta-reasoning system would have been designed to perform.

Section 2 considers these predictions in the light of psychological evidence of meta-reasoning abilities in humans. What the data show is that a disposition to reflect on one's reasoning is highly contingent on features of individual personality, and that the control of reflective reasoning is heavily dependent on learning, and especially on explicit training in norms and procedures for reasoning. In addition, people exhibit widely varied abilities to manage their own decision-making, employing a range of idiosyncratic techniques. These data count powerfully against the claim that humans possess anything resembling a system designed for reflecting on their own reasoning and decision-making. Instead, they support a view of meta-reasoning abilities as a diverse hodge-podge of self-management strategies acquired through individual and cultural learning, which co-opt whatever cognitive resources are available to serve monitoring-and-control functions.

Sections 3 and 4 outline a positive account of the architecture of human meta-reasoning, by considering the cognitive mechanisms that are recruited to implement self-management strategies. Section 3 lays out some groundwork, defending a particular model of system 2 as a *virtual* system realized in cycles of system 1 processing, mediated at each iteration by the rehearsal and manipulation of sensory imagery in working memory. On this view, system 2 is best

viewed, not as a distinctive reasoning 'system' with its own principles of operation, but rather as a programmable general architecture, which can support various culturally acquired routines and learned habits of reasoning. Section 4 elaborates this model into an account of meta-reasoning, which is explained in terms of the strategic recruitment of this imagery-rehearsal architecture for the purpose of self-reflection. Self-management strategies are informed by the self-directed use of a mindreading system, whose primary function is to interpret the minds of other agents, but which is here enlisted to monitor one's own reasoning processes on the basis of available sensory cues. This information is then used to control reasoning by exploiting the capacity for generating and manipulating sensory representations that forms a key part of system 2.

Two points should be stressed before we proceed. One is that we understand the 'meta' component of metacognition and meta-reasoning to require metarepresentation, or representations of one's own first-order mental states. This is fully in line with the psychological literature on metacognition, where the notion is defined as 'thinking about (one's own) thinking' [9,43], and therefore as involving meta-representation. The second point is that while metacognition always involves some degree of monitoring and control of a first-order process, there are many forms of monitoring and control that do not involve metarepresentation, and which therefore are not metacognitive in nature. Consider, for example, the use of forward models in the control of action. When motor schemata are activated and sent to the motor system to initiate an action, an efference copy of those instructions is sent to an emulator system that constructs a forward model of the expected sensory consequences of the movement [44–46]. This is received as input by a comparator mechanism that also receives reafferent sensory feedback, issuing in swift online adjustments in the action when there is a mismatch.

Note that the comparator system mentioned in this account is entirely non-metarepresentational in nature: it receives a sensory-coded representation of the intended outcome and compares this with sensory input from the actual outcome as it unfolds. When these fail to correspond, it employs an algorithm that adjusts the motor instructions to bring about a closer match. It does not need to represent, as such, either the motor intention or the current experiences resulting from the action. No metarepresentations are needed, and no one in the field of motor processing thinks that they are employed.

2. VARIABILITY IN META-REASONING

This section considers psychological evidence concerning people's ability to metarepresent and control their own reasoning and decision-making. The main finding is significant variation in meta-reasoning abilities and in the kinds of self-management techniques people employ. Indeed, many individuals seem not to reflect on their reasoning very much at all, while others appear to do so largely as a result of explicit cultural training (such as courses in mathematics, logic or scientific method). This is not what we would expect if there were a distinctive adaptation for meta-reasoning. Rather, these data support a view of meta-reasoning competence as residing in a diverse range of self-management habits, acquired through individual and cultural learning. We consider evidence of variability in reflective reasoning, in affect-based decision-making, and in self-control strategies.

(a) *Variability in reflective reasoning*

As we noted in §1, one of the principal functions attributed to system 2 is to override system 1, switching to the sort of controlled, reflective reasoning characteristic of system 2. Across a wide range of reasoning tasks, people appear to rely on system 1 by default, which swiftly gives rise to a response that intuitively 'seems right' to them. But one of the persistent findings of the heuristics-and-biases research programme is that these intuitive responses often fall short of normative standards of rationality [30,47]. It seems that the computational shortcuts employed by system 1 are apt to be led astray by misleading cues, resulting in sub-optimal judgements and decisions, at least as evaluated within the sorts of tasks that have been employed within this paradigm. These are precisely the sorts of situations, then, in which we would expect a dedicated monitoring-and-control system to override system 1 reasoning, frequently effecting a transition to slower, reflective, system 2 processing.

What the evidence shows, however, is significant variation in individuals' inclination to bring system 2 processing to bear in reasoning and decision-making [34,48]. Most subjects seem to rely exclusively on system 1 to perform these tasks, immediately endorsing, without further reflection, whatever response initially seems correct. Not surprisingly, these individuals are particularly prone to the sorts of biases and fallacies commonly associated with system 1. Other subjects, however, are more likely to switch to a reflective mode of reasoning, and thereby arrive more readily at normatively correct responses. Moreover, as some have emphasized, the difference between these two groups cannot be fully explained in terms of differences in 'algorithmic' features of their psychology, such as fluid intelligence or processing speed [35]. Rather, the feature primarily responsible for the superior performance of the second group is best understood as a personality trait they possess of being more 'reflective'—that is, being more disposed to pause and consider the adequacy of their initial reaction before settling on a final response.

This individual-differences finding would certainly be puzzling if human beings in general were equipped with special-purpose machinery for overriding system 1 by engaging in reflective, system 2, reasoning. For in that case, we would not expect the use of system 2 to be so fragile and contingent on features of individual personality. In contrast, these data make a good deal of sense if we regard reflective reasoning as, to a large extent, an acquired *habit*, which has been cultivated more successfully in some individuals than in others.

Another finding that is anomalous for the specific-adaptation proposal is the lack of natural competence people exhibit in evaluating their own system 2

reasoning. For as we pointed out in §1, a meta-reasoning system should be expected not only to effect transitions from an intuitive (system 1) to a reflective (system 2) mode of processing, but also to monitor and control system 2 itself, by noticing and correcting errors in one's subsequent reflective reasoning. Indeed, system 2 would seem to be a particularly good candidate for such metacognitive management, because its internal operations are not only subject to intentional control, but are also partly conscious and therefore accessible to the subject. (System 1 processes, on the other hand, function as 'black boxes'. Subjects have conscious access only to their sensory inputs and to the intuitive responses they produce as output.) So it is here that we might expect to see the signature effects of a meta-reasoning adaptation.

What the evidence shows, however, is that facility in evaluating one's own reasoning is a late-developing skill, emerging in late childhood or early adolescence, and then perhaps only as a result of formal instruction [49,50]. This suggests that in order to become competent at monitoring their own reflective reasoning, people must first acquire knowledge of the relevant reasoning norms, and only then are they able to notice shortcomings in their reasoning. Support for this interpretation is provided by a study demonstrating a close association between adolescents' explicit knowledge of various informal fallacies and their ability to detect instances of such errors in examples of reasoning [51]. It appears, then, that normatively informed management of one's own reasoning may be, to a large extent, a learned skill, acquired from one's culture through explicit instruction, rather than anything resembling an innate endowment. Indeed, the primary determinant of success in reasoning tasks that are tackled through the use of system 2 forms of reflection appears to be the adequacy of subjects' 'mindware'—that is to say, the explicit principles of reasoning they have acquired from the surrounding culture [35].

(b) *Variability in decision-making*

When confronted with difficult decisions, people will often imagine the various possible outcomes, respond affectively to each and choose the outcome that 'feels right' [52,53]. But some people will go further, making lists of positive and negative features, perhaps, as well as assigning weights to their importance and evaluating the probability of each outcome. In doing so, they may be relying on beliefs about norms of decision-making acquired from a class in economics, or they might be relying on the folk belief that one should carefully consider all the 'pros and cons' before taking an important decision. It turns out that in this domain, however, folk beliefs may be in error, and that people would be better off 'going with their gut' when confronted with a complex decision problem. For when subjects are encouraged to reflect on the 'good-making' and 'bad-making' features in a complex choice, they make normatively worse decisions than when they are prevented from engaging in such reflection [54–58]. It seems that the affective systems that make up part of system 1 can function together as an 'accumulator' that aggregates the various considerations, balancing the pros and cons off against each other in a way that is more optimal than proceeding by way of explicit, reflective, consideration.

In these and other similar decision-making tasks, the most effective self-management strategy will involve, not displacement of intuitive by reflective reasoning, but rather the use of reflection to modulate and redirect one's intuitive affective reactions. One example of such metacognitive management of affective reasoning is the discounting of incidental affect. There is a good deal of evidence that the affective responses that serve to inform our intuitive decision-making are often influenced by extraneous factors [59–61]. In such cases, affective feelings that are caused by one source are directed at another, quite different, target. For instance, someone who is angry following a frustrating day at work may, as a result, decide to lash out at his family when he returns home at the end of the day. Such arbitrary projections of emotion appear to stem from a default tendency for affect to be directed at whatever happens to occupy the current focus of attention. But this default can be overridden.

In a famous study, subjects were telephoned and asked to rate their overall life satisfaction [62]. This presented them with a simple decision-making task: to decide how positive or negative they should make their response. Not surprisingly, given the susceptibility of people to borrow affect from incidental sources, the local weather had an influence on their answers, with sunny weather being associated with relatively higher ratings, and gloomy weather with lower ratings. This effect disappeared, however, in trials where the conversation began with the interviewer casually asking, 'Thanks for agreeing to talk to us. How's the weather down there?' It appears that subjects who were reminded of the weather realized that their feelings of life satisfaction are prone to be influenced by the weather, and were able to discount that influence when deciding how strongly to respond.

This finding might already be thought to pose an anomaly for the idea of a specific meta-reasoning adaptation. For note that the subjects needed to be *reminded* that the weather is a potential source of some of their affective feelings before they would discount its influence. If there were really a meta-reasoning system designed for self-monitoring, one would expect subjects to be constantly alert to the possible sources of the affect they experience. On the other hand, once the incidental source was made salient, subjects appeared to be effective in engaging in appropriate self-management. However, this may be, in large part, a result of the heavy cultural reinforcement of 'common knowledge' concerning the influence of weather on mood. For in another similar study [63], in which good versus bad moods were induced by alternative means, there were significant individual differences in the efficacy of self-management, as we now explain.

Subjects were instructed to write about a recent event that was either happy or sad, and were then asked (in what was ostensibly a separate study) to provide various assessments of risk. (Note that the latter, again, is a kind of decision-making task, requiring

subjects to decide how highly to rate each estimate of risk; and successful performance requires subjects to monitor the influence of incidental affect on their decision-making. Moreover, the findings would presumably generalize to cases where subjects are required to decide whether or not to perform a risky action.) Writing about a happy rather than sad event was associated with lower risk estimates, as the induced affect was presumably brought to bear in determining the degree of risk. When the potential influence of recent events on mood was made salient to subjects, however (by asking them to write about events at that stage in the semester that might have had an impact on their mood), the effect that this had on assessments of risk depended on whether or not the subjects frequently attended to their own feelings. Those who did were able to discount the incidental affect induced at the outset of the study, and their risk assessments returned to normal. Those who did not often pay attention to their feelings, on the other hand, now displayed an *increased* effect of incidental mood on risk evaluations (presumably because for these subjects, attending to their feelings made them seem more relevant).

It might be suggested that a more neutral interpretation of these results is that those with little natural facility in discounting incidental affect have less incentive to pay attention to their feelings, because they are less able to put the information gained from such self-monitoring to effective use. However, this would fail to explain why subjects in general are able to discount affective influences that are acknowledged by prevailing folk wisdom, such as the weather. A better explanation is that those who were disposed to attend to their feelings had come to acquire, through learning, more effective techniques for discounting incidental affect. This provides further support for the picture that has been emerging of meta-reasoning as an acquired skill, rooted in learned strategies. (Compare the 'radical plasticity hypothesis' proposed in Timmermans *et al.* [64].)

(c) *Variability in self-control*

In some instances where system 1 affective processing pushes us rather directly towards a decision that would conflict with our interests, meta-reasoning can operate, not to modulate or transform, but rather to *suppress* our affective responses. Situations of temptation provide examples. When we are confronted with a temptation, affective feelings issuing from system 1 will give rise to a strong default tendency to act to secure whatever we are tempted by, even if doing so would be against our prudential interests or moral beliefs. The metacognitive task is to monitor our decision-making, intervening when an imminent decision is likely to be maladaptive. In such situations, the rationally optimal reaction is to resist temptation, which can be achieved by suspending or neutralizing the affective processing. Notice that here, metacognitive interventions do not lead to any further *reasoning*, but rather serve to interrupt or negate certain counterproductive affective responses. Nevertheless, insofar as these affective processes tend to issue in decisions to act, and are in that sense processes of decision-making, the regulatory processes that target them fall within the purview of meta-reasoning.

There is an extensive psychological literature on affective self-control [65–68]. Here, as with the kinds of reflective reasoning and affective decision-making we considered in §2*a*,*b*, there is ample evidence of significant individual differences in people's effectiveness in resisting temptation. Indeed, there is even evidence that the depletion that is often seen in people's willpower following attentionally demanding tasks depends on their folk beliefs about whether or not willpower is a limited, exhaustible resource [69].

What we particularly wish to stress, however, is the wide diversity of self-management *strategies* people employ in attempting to overcome situations of temptation. One such strategy involves pretending that a tempting object is merely a pictorial representation, by mentally 'putting a frame around it' [70]. By reframing the perceptual representation of the tempting stimulus in this manner, the 'hot' affective response that pushes one towards succumbing becomes significantly attenuated. A somewhat different strategy involves focusing on properties of the tempting object that are more abstract and less emotionally charged. As yet another alternative, people may intentionally distract their attention away from the object of their temptation by occupying themselves with unrelated real or imaginative activities. (Relatedly, some people employ the strategy of 'counting to ten' when angry, thus giving their initial 'hot' response a chance to cool, and reducing the chance of acting in ways that they might later regret.)

The significance of all this for our purposes is that such a diverse range of idiosyncratic strategies for managing temptation does not fit the profile of a distinctive competence for regulating one's own decision-making. If a specific adaptation really served to underpin people's self-control abilities, then we would expect them to reliably use the machinery that was designed for that purpose. Instead, self-control seems to operate by employing whatever imperfect self-regulatory techniques can be picked up from culture or from individual learning.

(d) *An adaptation for argumentation*

It appears that there are large individual differences in the extent to which people attempt to monitor and control their own reasoning and decision-making. It also appears that people possess little in the way of native competence for so doing, but instead are heavily dependent on individual and cultural learning. Moreover, in some contexts, their attempts to intervene are more of a hindrance than a help.

It might seem that these conclusions are inconsistent with the recently defended claim that humans possess an adaptation for monitoring, critiquing and improving argumentative inferences [71,72]. But there is no inconsistency. For the claimed adaptation is for monitoring and engaging in arguments with other people, not for monitoring and controlling one's own reasoning and decision-making. Indeed, much of the evidence cited in support of the account concerns contrasts between people's poor performance in reasoning tasks conducted alone with their abilities when engaged in argument or discussion with others. For example, if one group of subjects attempts conditional reasoning tasks individually, then only a

small number (9 per cent) may succeed. But if another set of subjects attempt the same tasks in small groups, then a massive 70 per cent may succeed [73]. It seems that people not only have the capacity to recognize correct solutions when proposed by others, but that such solutions can emerge out of processes of critical discussion [74,75]. Indeed, when groups are formed using only subjects who have previously failed at the tasks on an individual basis, 30 per cent of those groups can come up with the correct solution [73].

It might well be possible to co-opt the argumentation system in the service of private reasoning of a system 2 sort by providing oneself with appropriate interpersonal cues. For example, one might approach a reasoning or decision-making problem by imagining oneself as engaged in debate with others on the topic, playing through the debate in one's imagination and representing to oneself moves and counter-moves. But even if such a strategy can be successful, it will be just another contingent and variable tactic for monitoring and controlling one's own reasoning, on a par with the others discussed earlier. There is nothing here to challenge our account of meta-reasoning as a variable and cobbled-together set of habits and dispositions.

3. AN IMAGERY-REHEARSAL ARCHITECTURE FOR SYSTEM 2

This section lays some groundwork for the account of meta-reasoning to be outlined in §4. As we have seen in §2, human meta-reasoning abilities appear to reside, not in any distinctive native competence for metacognitive monitoring-and-control, but rather in a range of learned strategies of self-management that are implemented using the resources of system 2. Accordingly, we now turn our attention to the cognitive architecture underlying these system 2 processes. On the model defended here, which draws on the 'action-based' account provided by Carruthers [76], system 2 is not really a distinctive 'system' with its own principles of operation. It is rather a *virtual* system realized in cycles of imagery-rehearsal, which can be programmed to implement various learned habits and culturally acquired routines for reasoning [76–79]. These ideas build on those of Dennett [80], who suggests that the conscious mind is a sort of 'virtual machine' implemented in lower-level cognitive systems.

Recall that within the widely accepted dual-systems framework, system 1 is thought to comprise a *set* of intuitive systems arranged in parallel as consumers of perceptual information, which operate in a manner that is swift, unconscious, automatic and significantly innately specified. System 2, on the other hand, is held to be reflective in nature, characterized by serial, slow, conscious operations, which are subject to deliberate control and influenced by cultural learning—and in particular, by normative beliefs about how one should reason [35]. Moreover, system 2 is thought to be specific to humans, having evolved quite recently [34], with one of its principal functions being to override the evolutionarily more ancient system 1 processes.

This clustering of features gives rise to several puzzles regarding the cognitive architecture underlying system 2, as well as the sort of evolutionary history that might have produced it. On the assumption that the two systems are realized in distinct neural networks, for instance, it is not obvious how system 2 could be guided by verbal instructions. For this would suggest that such instructions somehow have the power to reach into the computational 'innards' of the system, altering their mode of operation. (In contrast, it is *actions*—taken in a broad sense to include allocations of attention—that can paradigmatically be guided by instruction.) It is similarly mysterious what evolutionary pressures might have driven selection for an entirely new reasoning system (rather than modifications to the existing one), given that system 2 duplicates much of the functionality of system 1. And yet another puzzle concerns the fact that system 2 reasoning is subject to intentional control and is influenced by culturally acquired norms about how one ought to reason. For active control is usually thought of as an *outcome* of reasoning, rather than something that serves to *guide* or *direct* reasoning processes.

These puzzles disappear, however, if system 2 is regarded, not as a neurally distinct system with its own dedicated hardware, but rather as a higher order *virtual* system, which is partly realized in cycles of system 1 activity [76–79]. On this sort of model, system 2 is capable of co-opting the reasoning resources belonging to system 1, making use of other available mechanisms in order to direct and structure the system 1 cycles. One such mechanism is a general architecture that situates the various system 1 modules as consumers of 'globally broadcast' [81] perceptual representations (including endogenously generated sensory imagery). Another is a set of emulator systems of the sort described in §1 [82]. In their evolutionarily original functional role, these use efferent copies of motor instructions to model the likely perceptual consequences of an action by generating sensory imagery, which can then be compared with the actual perceptual feedback in real time, facilitating online control of action [44–46]. These systems can be co-opted and used 'off-line' when motor plans are rehearsed, issuing primarily in visual, auditory or proprioceptive representations of the intended outcomes, but with motor instructions to the muscles suppressed. And when these emulator systems work in concert with language-production systems, the articulatory instructions issuing from the latter are transformed into auditory imagery in the so-called inner speech. This imagery is conceptualized and interpreted by language-comprehension and mind-reading systems while being globally broadcast to the full range of system 1 modules, which go to work interpreting it further, drawing relevant inferences, and issuing in appropriate affective responses.

This account can explain the main properties of system 2, while also avoiding the puzzles about the latter's existence raised earlier. Because globally broadcast images are conscious, this element in each cycle of mental rehearsal will also be conscious (while the cognitive activity that immediately precedes and follows the broadcast image will generally be *un*conscious). And because mental rehearsal activates and co-opts the resources of the various intuitive reasoning systems,

its overall operations are likely to be significantly slower than most of the latter. Nor is there any special difficulty in explaining how reflective reasoning could have evolved. For rather than existing alongside of intuitive reasoning systems while performing many of the same functions, reflection is partly realized in cycles of operation of the latter, using pre-existing mechanisms and capacities. All that had to evolve was a language system (for purposes of communication), together with a disposition to engage in mental rehearsal of action.

Moreover, because action selection in general is under intentional control and can be influenced by normative belief and verbal instruction, so can the operations of the described reflective system. We can *choose* (often unconsciously) to engage in mental rehearsal, just as we choose to engage in any other form of action. And just as with other forms of action, some sequences of rehearsal can be produced smoothly and automatically, resulting from previous practice. (Think, here, of doing a simple addition sum in your head.) Others can be guided by beliefs about how one *should* reason, sometimes by activating a stored memory of a previous instruction. (When faced with a conditional reasoning task, for example, one might rehearse the sentence, 'In order to evaluate a conditional, I should look for cases where the antecedent is true and the consequent false', or one might form a mental picture of the standard truth-table for the conditional.) And of course, with each iteration of mentally rehearsed action, the various system 1 systems that consume the globally broadcast images become active, sometimes producing an output that contains or contributes towards a solution to the problem in hand.

The account of system 2 just sketched has been expressed in terms of action in a fairly literal sense (involving the use of premotor and motor cortex; see [83–85]). In contrast, allocations of attention of the sort that are involved in intentionally calling to mind an item of visual imagery, or in framing or attending to a certain aspect of an image, are also important for the operations of system 2. While these are not literally actions, perhaps, they are at least active in the weaker sense of being subject to intentional control. All of these different elements of system 2, however, converge on the activation, maintenance, rehearsal and manipulation of various forms of imagery.

While this account of system 2 is, to some degree, controversial, it coheres nicely with what many in the field now regard as the defining feature of system 2. This is that the latter makes important use of the central-process working-memory system, whereas system 1 does not [35,39,48,86]. For example, system 2 processes tend to collapse under concurrent working-memory load, whereas system 1 processes do not [87]. Moreover, working memory itself seems to be constituted by attentionally controlled activation and active manipulation of sensory-involving representations [88–91]. These representations can be described as sensory-*involving* rather than merely sensory, however, because they often contain conceptual information bound into the content of the sensory images [92], in much the same way that conceptual content is bound into the content of perception [93]. Because these representations are globally broadcast to the full range of cognitive and affective systems designed to consume perceptual outputs (that is to say, to system 1), the upshot will be that controlled uses of working memory will be conscious, and will make their outputs available to system 1, just as we envisage earlier.

4. A SENSORY-BASED ACCOUNT OF META-REASONING

Section 2 argued that the available psychological evidence counts decidedly against the hypothesis of a specific meta-reasoning adaptation, and instead supports a view that sees human meta-reasoning competence as residing in a variety of habits and strategies for self-management, acquired by individual and cultural learning. For example, we noted in §2*c* that a common strategy for handling situations in which one finds oneself suddenly angry is to 'count to ten' before responding. Those who successfully internalize this strategy will presumably end up cultivating a receptivity to their own feelings of anger, noticing which will automatically cue the counting-to-ten routine into action, thereby forestalling an imprudent affect-driven response. Such acquired routines are of precisely the sort that we would expect to be handled by system 2. Accordingly, §3 was devoted to defending a view of the cognitive architecture underlying system 2 processes in general. The picture that emerged was not that of a distinctive reasoning system with fixed principles of operation, but rather of a programmable general architecture, based in cycles of imagery-rehearsal, capable of supporting a variety of culturally acquired reasoning routines.

The present section elaborates this model of system 2 into an account of meta-reasoning, by explaining how the various self-management routines co-opt available cognitive resources to serve metacognitive monitoring-and-control functions in the absence of any dedicated monitoring-and-control machinery. Because both the 'monitoring' and 'control' functions, on this model, are subserved by systems that traffic in sensory-involving representations, the upshot is a sensory-based account of meta-reasoning. But these functions are realized rather differently in cases where what is monitored-and-controlled are reasoning processes located in system 1, on the one hand, or in system 2, on the other. We therefore discuss these separately.

(a) *Monitoring of system 1 reasoning*

Because the internal processing of system 1 systems is unconscious, it is not available for metacognitive monitoring. As a result, monitoring of system 1 will be limited to its inputs and outputs, together with any indirect consequences of system 1 processing such as feelings of confidence or disfluency. The question for us is what it is about these that can trigger a shift to reflective modes of reasoning (or, as previously discussed, to a modulation or suppression of system 1 intuitive reasoning). Our view is that the relevant factors are likely to be quite heterogeneous and will vary widely among subjects. But all will be broadly sensory in character,

involving either perception of the environment or awareness of properties of the internal sensory milieu.

On the input side, some experienced subjects may know, for example, that certain kinds of reasoning problem require system 2 resources, and they may habitually activate such resources when problems of those kinds are identified. Thus, someone who has had frequent exposure to conditional reasoning tasks, and who has found it helpful in the past to consult the standard truth-table for the conditional, might now routinely activate a visual image of the truth-table as soon as a conditional reasoning task is recognized as such. Generalizing from this example, subjects may monitor the input to system 1 so as to initiate a switch to system 2 processing when responding to certain classes of problem (where the classes of problem in question will presumably vary widely among individuals, depending on their experience and training).

On the output side, there are a variety of cues that people can employ to trigger system 2 reasoning, or that might serve to suspend or redirect system 1 activity. Some people might have formed the habit, for example, of pausing to reflect whenever initial intuitions are accompanied by strongly felt affect, or by certain classes of affect. The 'count to ten' heuristic mentioned earlier as a response to anger might be an instance of such a habit. Another common strategy is to switch to system 2 processing whenever the output of system 1 is accompanied by feelings of *disfluency* [94], or whenever the intuitive answers produced by system 1 are accompanied by feelings of anxiety or uncertainty. But not everyone employs such strategies, of course. (If they did, then we would expect to see much more reflective reasoning taking place in standard reasoning tasks than we actually observe.) And the strategies employed are likely to vary across individuals, depending on accidents of temperament and learning history.

It might be objected that people can monitor and respond to their own cognitive processes more swiftly than system 1 systems can produce a response [95–97]. This might be thought to support the existence of a specific metacognitive mechanism designed for the purpose. For example, when deciding whether or not the answer to a math problem can be retrieved from memory or needs to be calculated on a scrap of paper, results show that people can estimate whether the answer is known faster than the answer can be retrieved [96]. But the mechanism thought to be responsible for this ability is prior familiarity of the problem, which elicits a feeling of knowing. Notice, therefore, that it is not metacognitive monitoring that produces these feelings. Rather, such monitoring, if it occurs, will occur downstream, and be monitoring *of* the feeling of knowing. Moreover, such feelings may not even need to be monitored and classified as such (presumably by the mindreading faculty) in order to achieve their effects. Rather, because we know that familiarity produces positive affect [98,99], it may be that positive affect directed towards the familiar math problem is sufficient to motivate attempts to recall the answer without the intervention of any metacognitive monitoring.

Indeed, a case can be made that, in general, epistemic emotions such as confidence or uncertainty achieve their effects in a first-order way, without involving metacognitive resources [10]. On this account, the exercise of executive control or the use of the 'supervisory attentional system' [100] may frequently require only first-order, non-metarepresentational resources. And it is important to notice, indeed, that only some of the forms of monitoring discussed here require such resources. (There is a contrast in this respect with system 2 reasoning, as we will see in §4*c*.) Monitoring of the circumstances that provide inputs for system 1, for example ('this is a conditional task'), can be entirely first order. And monitoring of disfluency can involve first-order cues such as the elapse of time between presentation of the task and the resulting system 1 intuition. Only monitoring and classifying emotional cues as such (to the extent that this occurs) is likely to require metarepresentation. It follows, then, that many of the factors that initiate a switch from system 1 to system 2 processing are not strictly metacognitive in nature.

(b) *Control of system 1 reasoning*

System 1 reasoning cannot be controlled directly, of course. But its output can be ignored or suppressed, and its input can be manipulated. The former may occur whenever someone relies upon situational cues (such as the nature of the reasoning task presented), properties of the output of system 1 (such as strong affect) or properties accompanying the output of system 1 (such as uncertainty or disfluency) to initiate system 2 forms of reasoning. In contrast, the latter may occur whenever subjects respond to cues of these sorts by attempting to manipulate the input to system 1. The 'count to ten' heuristic would be one example of this. For by switching one's attention to a simple behavioural task, one thereby distracts attention from the anger-provoking stimulus and provides time for the anger generated by system 1 affective systems to dissipate. Other examples are the 'framing' heuristics and distraction techniques discussed in §2*c*.

(c) *Monitoring of system 2 reasoning*

System 2 reasoning comprises at least two rather different varieties. One is discursive, and often involves the rehearsal of representations of natural language phrases and sentences (generally in the form of auditory imagery). But it might also incorporate visual imagery, such as a visual representation of the standard truth-table for the conditional. The other (less well studied) form of system 2 reasoning involves imagination and appraisal. It will often include the activation and manipulation of visual images, responding affectively to each. But it can also involve imagining natural language utterances. For instance, one might imagine how someone would react if one were to say one thing rather than another, responding affectively to the imagined outcome. (The difference between this and discursive reasoning is that here one is reasoning about *what to say*, not the subject matter of what one says.)

Consider imagination–appraisal reasoning first. We know that people employ imagistic representations of actions or outcomes when reasoning about what they should do [52,53]. Sometimes these images might be

activated from memory or generated from affordances of the situation, but they also frequently involve activations of the motor schemata for some of the actions that are open to one in the circumstances, issuing in a sensory 'forward model' of the likely outcomes of those actions. These images are globally broadcast in the manner of perception and are received as input by all system 1 systems, including one's evaluative and affective systems. The latter respond with some degree of affect, which is likewise globally broadcast and monitored by the subject. As a result, one's motivations towards the represented actions or outcomes are adjusted up or down, sometimes issuing in a decision.

Because the form of reasoning described here is at least minimally reflective and involves the controlled use of working memory, it might seem that it already qualifies as system 2, even if its operations are subject to a variety of well-known fallacies and biases [53]. As described so far, however, it merely involves the iterated use of system 1 affective systems responding to imagined possibilities, issuing in intuitions of desirability. Hence, simple forms of imagination–appraisal reasoning are no more system 2 than are one's immediate responses to a question about probabilities or about the evaluation of a conditional. Both will involve the initial activation of representations in working memory, and each will issue in an intuition. Nevertheless, some subjects may employ more elaborate strategies for engaging in prospective reasoning. For example, they might imagine not just the events themselves but also their surrounding circumstances or their likely consequences [53], responding affectively to these also.

What we suggest is that imagination–appraisal reasoning qualifies as system 2 to the extent that involves more than minimal executive management of the sequence of imaginings, including not only the imagining of consequences, but also reflecting on whether one is imagining realistically, making sure one has imagined all the available options, circling back on previously imagined scenarios to make sure one has a grip on comparative values across different options, and so forth.

Now consider discursive forms of system 2 reasoning, which often employ representations of natural language sentences in the so-called inner speech. This, too, involves the off-line activation of motor schemata (generally articulatory instructions for speech, in the case of hearing speakers, or motor instructions for hand movements, in the case of signers). These are used to generate globally broadcast representations of the speech acts that would result. Sequences of such sentences can be created by following well-worn speech habits or routines (as when one does an addition or multiplication sum in one's head). Alternatively, they can be guided either by system 1 intuitions or by one's beliefs about normative standards of reasoning. Indeed, because this form of system 2 reasoning is fundamentally *active* in nature, it can be guided and controlled in any of the ways that overt speech can be guided and controlled.

System 2 reasoning of both of the earlier mentioned varieties will generally implicate the metarepresentational resources of the mindreading faculty [10]. In the case of inner speech, the mindreading system will work together with the language comprehension system to issue in something that will be experienced as *wondering whether to go to the lake*, for example (just as often happens in connection with the speech of other people). And when rehearsing one of the actions that are open to one, the mindreading system will generally interpret the resulting images in such a way that one experiences oneself as *imagining* doing something (as opposed to remembering it or seeing it). For sensory representations in general need to be *interpreted* in order to be classified as involving one sort of mental attitude rather than another.

Before we discuss what is involved when we monitor our own system 2 reasoning, it is important to note that none of the resources from which system 2 is built are dedicated to the purpose. On the contrary, the primary function of the efference copies and emulator systems that produce forward models of activated motor schemata is the online control of movement [46], and the capacity to rehearse sentences in inner speech likewise rides on the back of such functions. Moreover, the language production and comprehension systems evolved for the purposes of communication, but can be co-opted for the creation and interpretation of self-generated sentences in inner speech. Finally, while the primary function of the mindreading system is to attribute mental states to other agents, it can be turned towards the self, using the same sensory channels and many of the same inferential principles that are employed for third-person mindreading [10].

Key components of system 2 reasoning are conscious, of course, by virtue of the global broadcast of the imagistic representations involved. But what sorts of cues are used to signal that system 2 reasoning is not proceeding correctly, and needs to be intervened in or redirected? In some instances, memories of culturally acquired information can be evoked by the reasoning process itself. For example, someone engaged in prospective reasoning using visual imagery might recall reading about some of the common fallacies involved, and attempt to correct for them—perhaps by attempting to imagine the surrounding spatial and temporal context of the represented outcome in addition to the outcome itself [53]. Similarly, someone engaged in verbal reasoning might recognize a fallacy in the sequence of sentences just entertained, and attempt to generate a sequence of sentences that avoids the fallacy.

If there is an adaptation for monitoring the arguments of others and identifying their strengths and weaknesses, as well as for generating persuasive arguments for oneself [71,72], then this mechanism may well play an important role in the operations of system 2 (at least in some people, on some occasions). In the first place, since it is hypothesized to play a role in the construction of arguments designed to convince others, the mechanism might sometimes be involved in the initial construction of system 2 reasoning, provided one adopts the strategy of arguing with an imagined opponent. For it seems likely that the mechanism can be evoked into activity by representations of an imagined as well as a real opponent. But the sentences thereby produced may also cue the mechanism into its critical mode. The cues that are monitored so as to

issue in corrections and improvements in system 2 reasoning will then include whatever cues the argumentation system uses when monitoring the arguments of other people.

(d) *Control of system 2 reasoning*

If the account that we have outlined of system 2 reasoning is correct, then it should already be plain how one can control the course of such reasoning when fallacies or errors are detected. Because system 2 is largely action-based, resulting from directed uses of attention and mental rehearsals of action-schemata of various sorts, it can be controlled and guided in any of the ways that overt action can. In particular, one can use one's beliefs about norms of reasoning to generate sequences of imagistic representations that conform to those norms. And one can likewise activate well-rehearsed behavioural routines, including habits of self-interrogation ('What should I do now?') and trained numeric and other sequences ('Seven sevens are forty nine'.)

(e) *Is there a meta-reasoning system?*

We have sketched some of the ways in which monitoring-and-control functions can be realized in a dual-systems reasoning architecture. But does this entitle us to speak of a distinctive *system* for meta-reasoning? The term 'system' can be interpreted in multiple ways, of course. One manner of making it more precise is to stipulate that a cognitive system should be universal among normal humans (think of working memory, or the language faculty, for examples). Another would be to stipulate that a cognitive system should exhibit a common structure among all normal individuals who possess it, even if it is not universal (the reading system might be a plausible instance of this).

It is doubtful whether system 2 constitutes a system in either of the earlier mentioned senses, although its various manifestations may draw on a common set of resources (including working memory, the forward modelling of action and the mindreading system). For system 2 reasoning can comprise a number of different processes, including visual imagery of potential actions and outcomes that are responded to affectively by the subject, as well as auditory imagery of sequences of sentences in inner speech. Moreover, introspection-sampling studies suggest that some individuals almost never employ inner speech (while others do so as much as 75 per cent of the time), whereas other people never employ visual imagery (while others do so as much as 90 per cent of the time) [101]. Likewise, tests of reasoning ability suggest that many people never draw on system 2 resources to solve the problems with which they are presented [34,48]. It seems likely, however, that all groups possess the *capacity* for both visual imagery and inner speech; for these require only working memory and the capacity for forward modelling of action, which are surely universal.

So far, these points are consistent with the idea that system 2 comprises a number of distinct systems of reasoning, the frequency of whose use varies among individuals, but which possess a common structure in all people who use them. Recall, however, that the ways in which individuals attempt to tackle reasoning problems are quite varied, even when approached in some particular system 2 mode, drawing on a range of both culture-dependent and idiosyncratic habits and normative beliefs. There is little to suggest the existence of a common processing structure among all individuals who employ inner speech, for example; and hence little to suggest the presence of a *system* for system 2 reasoning in even the weaker of the two senses distinguished earlier.

It is even more doubtful whether there is a system for meta-reasoning. Granted, meta-reasoning draws on the same set of basic resources that realize system 2, in general. These include global broadcasting of conceptualized sensory representations (on the 'monitoring' side), the capacity for controlled rehearsal of action and self-produced imagery in working memory (on the 'control' side), together with a set of learned habits and beliefs about norms of reasoning that serve to map the former onto the latter in various specific ways. But the only components that are specifically concerned with meta-reasoning are these habits and beliefs themselves. Moreover, these are idiosyncratic and acquired piecemeal. The result is that a diverse set of cues and self-management routines and principles are likely to be employed, varying widely by culture and by individual. The cues in question are all broadly sensory in character, however; and system 2 itself is partly constituted by sensory-involving representations, together with acquired beliefs, well-rehearsed motor schemata and more.

5. CONCLUSION

We have argued that there is not a distinct and distinctive metacognitive system for monitoring and controlling one's reasoning; let alone an innately channelled system designed for the purpose. Rather, widely varied kinds of cue can be used for monitoring one's reasoning of both system 1 and system 2 sorts (but all of which are broadly sensory in nature), varying by individual and by culture. Likewise, there are varied kinds of strategy for intervening in or directing one's reasoning, again varying by individual and culture.

We are grateful to Steve Fleming and Chris Frith for their advice and feedback on earlier drafts of this essay, and to an anonymous reviewer for a very helpful set of critical comments.

REFERENCES

1 Reder, L. 1987 Strategy selection in question answering. *Cogn. Psychol.* **19**, 90–138. (doi:10.1016/0010-0285(87)90005-3)
2 Begg, I., Duft, S., Lalonde, P., Melnick, R. & Sanvito, J. 1989 Memory predictions are based on ease of processing. *J. Mem. Lang.* **28**, 610–632. (doi:10.1016/0749-596X(89)90016-8)
3 Leonesio, R. & Nelson, T. 1990 Do different metamemory judgments tap the same underlying aspects of memory? *J. Exp. Psychol. Learn.* **16**, 464–470. (doi:10.1037/0278-7393.16.3.464)
4 Nelson, T. & Narens, L. 1990 Metamemory: a theoretical framework and new findings. In *The psychology of*

learning and information, vol. 26 (ed. G. Bower), pp. 125–173. San Diego, CA: Academic Press.
5 Metcalfe, J., Schwartz, B. & Joaquim, S. 1993 The cue-familiarity heuristic in metacognition. *J. Exp. Psychol. Learn.* **19**, 851–861. (doi:10.1037/0278-7393.19.4.851)
6 Koriat, A. 1993 How do we know that we know? The accessibility model of the feeling of knowing. *Psychol. Rev.* **100**, 609–639. (doi:10.1037/0033-295X.100.4.609)
7 Koriat, A. 1995 Dissociating knowing and the feeling of knowing: further evidence for the accessibility model. *J. Exp. Psychol. Gen.* **124**, 311–333. (doi:10.1037/0096-3445.124.3.311)
8 Koriat, A. 1997 Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *J. Exp. Psychol. Gen.* **126**, 349–370. (doi:10.1037/0096-3445.126.4.349)
9 Dunlosky, J. & Metcalfe, J. 2009 *Metacognition*. Los Angeles, CA: Sage Publications.
10 Carruthers, P. 2011 *The opacity of mind*. Oxford, UK: Oxford University Press.
11 Chua, E., Schacter, D., Rand-Giovannetti, E. & Sperling, R. 2006 Understanding metamemory: neural correlates of the cognitive process and subjective level of confidence in recognition memory. *NeuroImage* **29**, 1150–1160. (doi:10.1016/j.neuroimage.2005.09.058)
12 Chua, E., Schacter, D. & Sperling, R. 2009 Neural correlates of metamemory: a comparison of feeling-of-knowing and retrospective confidence judgments. *J. Cogn. Neurosci.* **21**, 1751–1765. (doi:10.1162/jocn.2009.21123)
13 Saxe, R. 2009 Theory of mind (neural basis). In *Encyclopedia of consciousness* (ed. W. Banks), pp. 401–410. Oxford, UK: Academic Press.
14 Fleming, S. M. & Dolan, R. J. 2012 The neural basis of metacognitive ability. *Phil. Trans. R. Soc. B* **367**, 1338–1349. (doi:10.1098/rstb.2011.0417)
15 Hampton, R. 2001 Rhesus monkeys know when they remember. *Proc. Natl Acad. Sci. USA* **98**, 5359–5362. (doi:10.1073/pnas.071600998)
16 Hampton, R., Zivin, A. & Murray, E. 2004 Rhesus monkeys (*Macaca mulatta*) discriminate between knowing and not knowing and collect information as needed before acting. *Anim. Cogn.* **7**, 239–246. (doi:10.1007/s10071-004-0215-1)
17 Smith, J., Shields, W. & Washburn, D. 2003 The comparative psychology of uncertainty monitoring and meta-cognition. *Behav. Brain Sci.* **26**, 317–373.
18 Smith, J., Beran, M., Redford, J. & Washburn, D. 2006 Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *J. Exp. Psychol. Gen.* **135**, 282–297. (doi:10.1037/0096-3445.135.2.282)
19 Smith, J., Redford, J., Beran, M. & Washburn, D. 2010 Rhesus monkeys (*Macaca mulatta*) adaptively monitor uncertainty while multi-tasking. *Anim. Cogn.* **13**, 93–101. (doi:10.1007/s10071-009-0249-5)
20 Beran, M., Smith, J., Coutinho, M., Couchman, J. & Boomer, J. 2009 The psychological organization of 'uncertainty' responses and 'middle' responses: a dissociation in capuchin monkeys (*Cebus apella*). *J. Exp. Psychol. Anim. B.* **35**, 371–381. (doi:10.1037/a0014626)
21 Couchman, J., Coutinho, M., Beran, M. & Smith, J. 2010 Beyond stimulus cues and reinforcement signals: a new approach to animal metacognition. *J. Comp. Psychol.* **124**, 356–368. (doi:10.1037/a0020129)
22 Washburn, D., Gulledge, J., Beran, M. & Smith, J. 2010 With his memory magnetically erased, a monkey knows he is uncertain. *Biol. Lett.* **6**, 160–162. (doi:10.1098/rsbl.2009.0737)
23 Smith, J. D., Couchman, J. J. & Beran, M. J. 2012 The highs and lows of theoretical interpretation in animal-metacognition research. *Phil. Trans. R. Soc. B* **367**, 1297–1309. (doi:10.1098/rstb.2011.0366)
24 Couchman, J., Coutinho, M., Beran, M. & Smith, D. 2009 Metacognition is prior. *Behav. Brain Sci.* **32**, 142. (doi:10.1017/S0140525X09000594)
25 Gallese, V. & Goldman, A. 1998 Mirror neurons and the simulation theory of mindreading. *Trends Cogn. Sci.* **2**, 493–501. (doi:10.1016/S1364-6613(98)01262-5)
26 Nichols, S. & Stich, S. 2003 *Mindreading*. Oxford, UK: Oxford University Press.
27 Goldman, A. 2006 *Simulating minds*. Oxford, UK: Oxford University Press.
28 Anderson, M. & Perlis, D. 2005 Logic, self-awareness and self-improvement. *J. Logic Comput.* **15**, 21–40. (doi:10.1093/logcom/exh034)
29 Rolls, E. 2007 *Memory, attention, and decision-making*. Oxford, UK: Oxford University Press.
30 Kahneman, D., Slovic, P. & Tversky, A. (eds) 1982 *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.
31 Evans, J. & Over, D. 1996 *Rationality and reasoning*. East Sussex, UK: Psychology Press.
32 Sloman, S. 1996 The empirical case for two systems of reasoning. *Psychol. Bull.* **119**, 3–22. (doi:10.1037/0033-2909.119.1.3)
33 Sloman, S. 2002 Two systems of reasoning. In *Heuristics and biases* (eds T. Gilovich, D. Griffin & D. Kahneman), pp. 379–396. Cambridge, UK: Cambridge University Press.
34 Stanovich, K. 1999 *Who is rational?* Mahwah, NJ: Erlbaum.
35 Stanovich, K. 2009 *What intelligence tests miss: the psychology of rational thought*. New Haven, CT: Yale University Press.
36 Kahneman, D. & Frederick, S. 2002 Representativeness revisited: attribute substitution in intuitive judgment. In *Heuristics and biases* (eds T. Gilovich, D. Griffin & D. Kahneman), pp. 49–81. Cambridge, UK: Cambridge University Press.
37 Kahneman, D. 2003 A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* **58**, 697–720. (doi:10.1037/0003-066X.58.9.697)
38 Kahneman, D. 2011 *Thinking, fast and slow.* New York, NY: Farrar, Straus and Giroux.
39 Evans, J. 2008 Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* **59**, 255–278. (doi:10.1146/annurev.psych.59.103006.093629)
40 Gigerenzer, G. & Regier, T. 1996 How do we tell an association from a rule? Comment on Sloman (1996). *Psychol. Bull.* **119**, 23–26. (doi:10.1037/0033-2909.119.1.23)
41 Gigerenzer, G., Todd, P. & the ABC Research Group. 1999 *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.
42 Kruglanski, A. & Gigerenzer, G. 2011 Intuitive and deliberate judgments are based on common principles. *Psychol. Rev.* **118**, 97–109. (doi:10.1037/a0020762)
43 Flavell, J. 1979 Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *Am. Psychol.* **34**, 906–911. (doi:10.1037/0003-066X.34.10.906)
44 Wolpert, D. & Kawato, M. 1998 Multiple paired forward and inverse models for motor control. *Neural Networks* **11**, 1317–1329. (doi:10.1016/S0893-6080(98)00066-5)
45 Wolpert, D. & Ghahramani, Z. 2000 Computational principles of movement neuroscience. *Nat. Neurosci.* **3**, 1212–1217. (doi:10.1038/81497)
46 Jeannerod, M. 2006 *Motor cognition*. Oxford, UK: Oxford University Press.

47 Tversky, A. & Kahneman, D. 1974 Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131. (doi:10.1126/science.185.4157.1124)
48 Stanovich, K. & West, R. 2000 Individual differences in reasoning: implications for the rationality debate. *Behav. Brain Sci.* **23**, 645–726. (doi:10.1017/S0140525X00003435)
49 Pillow, B. 2002 Children's and adults' evaluation of certainty of deductive inference, inductive inference, and guesses. *Child Dev.* **73**, 779–792. (doi:10.1111/1467-8624.00438)
50 Moshman, D. 2004 From inference to reasoning: the construction of rationality. *Think. Reason.* **10**, 221–239. (doi:10.1080/13546780442000024)
51 Weinstock, M., Neuman, Y. & Tabak, I. 2004 Missing the point or missing the norms? Epistemological norms as predictors of students' ability to identify fallacious arguments. *Contemp. Educ. Psychol.* **29**, 77–94. (doi:10.1016/S0361-476X(03)00024-9)
52 Damasio, A. 1994 *Descartes' error*. New York, NY: Putnam Publishing.
53 Gilbert, D. & Wilson, T. 2007 Prospection: experiencing the future. *Science* **317**, 1351–1354. (doi:10.1126/science.1144161)
54 Wilson, T., Dunn, D., Kraft, D. & Lisle, D. 1989 Introspection, attitude change, and attitude behavior consistency: the disruptive effects of explaining why we feel the way we do. In *Advances in experimental social psychology*, vol. 22 (ed. L. Berkowitz), pp. 287–343. San Diego, CA: Academic Press.
55 Wilson, T., Lisle, D., Schooler, J., Hodges, S., Klaaren, K. & LaFleur, S. 1993 Introspecting about reasons can reduce post-choice satisfaction. *Pers. Soc. Psychol. B.* **19**, 331–339. (doi:10.1177/0146167293193010)
56 Dijksterhuis, A. 2004 Think different: the merits of unconscious thought in preference development and decision making. *J. Pers. Soc. Psychol.* **87**, 586–598. (doi:10.1037/0022-3514.87.5.586)
57 Dijksterhuis, A., Bos, M., Nordgren, L. & van Baaren, R. 2006 On making the right choice: the deliberation-without-attention effect. *Science* **311**, 1005–1007. (doi:10.1126/science.1121629)
58 Mikels, J., Maglio, S., Reed, A. & Kaplowitz, L. 2011 Should I go with my gut? Investigating the benefits of emotion-focused decision making. *Emotion* **11**, 743–753. (doi:10.1037/a0023986)
59 Kunda, Z. 1999 *Social cognition*. Cambridge, MA: MIT Press.
60 Schwarz, N. & Clore, G. 2000 Mood as information: 20 years later. *Psychol. Inq.* **14**, 296–303.
61 Winkielman, P., Berridge, K. & Wilbarger, J. 2005 Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Pers. Soc. Psychol. B.* **31**, 121–135. (doi:10.1177/0146167204271309)
62 Schwarz, N. & Clore, G. 1983 Mood, misattribution, and judgments of well-being: informative affective states. *J. Pers. Soc. Psychol.* **45**, 513–523. (doi:10.1037/0022-3514.45.3.513)
63 Gasper, K. & Clore, G. 2000 Do you have to pay attention to your feelings to be influenced by them? *Pers. Soc. Psychol. B.* **26**, 698–711. (doi:10.1177/0146167200268005)
64 Timmermans, B., Schilbach, L., Pasquali, A. & Cleeremans, A. 2012 Higher-order thoughts in action: consciousness as an unconscious redescription process. *Phil. Trans. R. Soc. B* **367**, 1412–1423. (doi:10.1098/rstb.2011.0421)
65 Shoda, Y., Mischel, W. & Peake, P. 1990 Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: identifying diagnostic conditions. *Dev. Psychol.* **26**, 978–986. (doi:10.1037/0012-1649.26.6.978)
66 Metcalfe, J. & Mischel, W. 1999 A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychol. Rev.* **106**, 3–19. (doi:10.1037/0033-295X.106.1.3)
67 Masicampo, E. & Baumeister, R. 2008 Toward a physiology of dual-process reasoning and judgment: lemonade, willpower, and expensive rule-based analysis. *Psychol. Sci.* **19**, 255–260. (doi:10.1111/j.1467-9280.2008.02077.x)
68 Mischel, W. & Ayduk, O. 2004 Willpower in a cognitive affective processing system: the dynamics of delay of gratification. In *Handbook of self-regulation* (eds R. Baumeister & K. Vohs), pp. 99–129. New York, NY: Guilford Press.
69 Job, V., Dweck, C. & Walton, G. 2010 Ego depletion—is it all in your head? Implicit theories about willpower affect self-regulation. *Psychol. Sci.* **21**, 1686–1693. (doi:10.1177/0956797610384745)
70 Mischel, H. & Mischel, W. 1983 The development of children's knowledge of self-control strategies. *Child Dev.* **54**, 603–619. (doi:10.2307/1130047)
71 Mercier, H. & Sperber, D. 2009 Intuitive and reflective inferences. In *In two minds* (eds J. Evans & K. Frankish), pp. 149–170. Oxford, UK: Oxford University Press.
72 Mercier, H. & Sperber, D. 2011 Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* **34**, 57–74. (doi:10.1017/S0140525X10000968)
73 Moshman, D. & Geil, M. 1998 Collaborative reasoning: evidence for collective rationality. *Think. Reason.* **4**, 231–248. (doi:10.1080/135467898394148)
74 Schulz-Hardt, S., Brodbeck, F., Mojzisch, A., Kerschreiter, R. & Frey, D. 2006 Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *J. Pers. Soc. Psychol.* **91**, 1080–1093. (doi:10.1037/0022-3514.91.6.1080)
75 Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G. & Frith, C. 2012 What failure in collective decision-making tells us about metacognition. *Phil. Trans. R. Soc. B* **367**, 1350–1365. (doi:10.1098/rstb.2011.0420)
76 Carruthers, P. 2009 An architecture for dual reasoning. In *In two minds* (eds J. Evans & K. Frankish), pp. 109–128. Oxford, UK: Oxford University Press.
77 Frankish, K. 2004 *Mind and supermind*. Cambridge, UK: Cambridge University Press.
78 Frankish, K. 2009 Systems and levels. In *In two minds* (eds J. Evans & K. Frankish), pp. 89–108. Oxford, UK: Oxford University Press.
79 Carruthers, P. 2006 *The architecture of the mind*. Oxford, UK: Oxford University Press.
80 Dennett, D. 1991 *Consciousness explained*. New York, NY: Little, Brown & Co.
81 Baars, B. 1988 *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.
82 Grush, R. 2004 The emulation theory of representation: motor control, imagery, and perception. *Behav. Brain Sci.* **27**, 377–442.
83 Paulescu, E., Frith, C. & Frackowiak, R. 1993 The neural correlates of the verbal component of working memory. *Nature* **362**, 342–345. (doi:10.1038/362342a0)
84 Richter, W. *et al.* 2000 Motor area activity during mental rotation studied by time-resolved single-trial fMRI. *J. Cogn. Neurosci.* **12**, 310–320. (doi:10.1162/089892900562129)
85 Lamm, C., Windtschberger, C., Leodolter, U., Moser, E. & Bauer, H. 2001 Evidence for premotor cortex activity during dynamic visuospatial imagery from single trial functional magnetic resonance imaging and

event-related slow cortical potentials. *Neuroimage* **14**, 268–283. (doi:10.1006/nimg.2001.0850)
86 Barrett, L., Tugade, M. & Engle, R. 2004 Individual differences in working memory capacity and dual-process theories of the mind. *Psychol. Bull.* **130**, 553–573. (doi:10.1037/0033-2909.130.4.553)
87 De Neys, W. 2006 Dual processing in reasoning: two systems but one reasoner. *Psychol. Sci.* **17**, 428–433. (doi:10.1111/j.1467-9280.2006.01723.x)
88 Müller, N. & Knight, R. 2006 The functional neuroanatomy of working memory: contributions of human brain lesion studies. *Neuroscience* **139**, 51–58. (doi:10.1016/j.neuroscience.2005.09.018)
89 Postle, B. 2006 Working memory as an emergent property of the mind and brain. *Neuroscience* **139**, 23–38. (doi:10.1016/j.neuroscience.2005.06.005)
90 D'Esposito, M. 2007 From cognitive to neural models of working memory. *Phil. Trans. R. Soc. B* **362**, 761–772. (doi:10.1098/rstb.2007.2086)
91 Jonides, J., Lewis, R., Nee, D., Lustig, C., Berman, M. & Moore, K. 2008 The mind and brain of short-term memory. *Annu. Rev. Psychol.* **59**, 193–224. (doi:10.1146/annurev.psych.59.103006.093615)
92 Baddeley, A. 2006 *Working memory, thought, and action*. Oxford, UK: Oxford University Press.
93 Kosslyn, S. 1994 *Image and brain*. Cambridge, MA: MIT Press.
94 Alter, A., Oppenheimer, D., Epley, N. & Eyre, R. 2007 Overcoming intuition: metacognitive difficulty activates analytic reasoning. *J. Exp. Psychol. Gen.* **136**, 569–576. (doi:10.1037/0096-3445.136.4.569)
95 Reder, L. & Ritter, F. 1992 What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *J. Exp. Psychol. Learn.* **18**, 435–451. (doi:10.1037/0278-7393.18.3.435)
96 Paynter, C., Reder, L. & Kieffaber, P. 2009 Knowing we know before we know: ERP correlates of initial feeling-of-knowing. *Neuropsychologia* **47**, 796–803. (doi:10.1016/j.neuropsychologia.2008.12.009)
97 Walsh, M. & Anderson, J. 2009 The strategic nature of changing your mind. *Cogn. Psychol.* **58**, 416–440. (doi:10.1016/j.cogpsych.2008.09.003)
98 Zajonc, R. 1968 The attitudinal effects of mere exposure. *J. Pers. Soc. Psychol.* **8**, 264–288.
99 Zajonc, R. 2001 Mere exposure: a gateway to the subliminal. *Curr. Dir. Psychol. Sci.* **10**, 224–229. (doi:10.1111/1467-8721.00154)
100 Norman, D. & Shallice, T. 2000 Attention to action: willed and automatic control of behavior. In *Cognitive neuroscience* (ed. M. Gazzaniga), pp. 376–390. Malden, MA: Blackwell.
101 Heavey, C. & Hurlburt, R. 2008 The phenomena of inner experience. *Conscious. Cogn.* **17**, 798–810. (doi:10.1016/j.concog.2007.12.006)