Biol Philos
DOI 10.1007/s10539-009-9171-0

Metamemory as evidence of animal consciousness: the type that does the trick

Nicholas Shea · Cecilia Heyes

Received: 25 September 2008 / Accepted: 18 May 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract The question of whether non-human animals are conscious is of fundamental importance. There are already good reasons to think that many are, based on evolutionary continuity and other considerations. However, the hypothesis is notoriously resistant to direct empirical test. Numerous studies have shown behaviour in animals analogous to consciously-produced human behaviour. Fewer probe whether the same mechanisms are in use. One promising line of evidence about consciousness in other animals derives from experiments on metamemory. A study by Hampton (Proc Natl Acad Sci USA 98(9):5359–5362, 2001) suggests that at least one rhesus macaque can use metamemory to predict whether it would itself succeed on a delayed matching-to-sample task. Since it is not plausible that mere meta-representation requires consciousness, Hampton's study invites an important question: what kind of metamemory *is* good evidence for consciousness? This paper argues that if it were found that an animal had a memory trace which allowed it to use information about a past perceptual stimulus to inform a range of different behaviours, that would indeed be good evidence that the animal was conscious. That functional characterisation can be tested by investigating whether successful performance on one metamemory task transfers to a range of new tasks. The paper goes on to argue that thinking about animal consciousness in this way helps in formulating a more precise functional characterisation of the mechanisms of conscious awareness.

N. Shea (✉)
Faculty of Philosophy, University of Oxford, 10 Merton Street, Oxford OX1 4JJ, UK
e-mail: nicholas.shea@philosophy.ox.ac.uk

N. Shea
Somerville College, Oxford, UK

C. Heyes
All Souls College, Oxford, UK

Published online: 29 July 2009

Keywords Animal consciousness · Metamemory · Phenomenal consciousness · Meta-representation · Higher order theories of consciousness · Global availability · Access consciousness

Investigating animal consciousness. Why? How?

There is a wealth of research on animals' metacognitive abilities. Some experiments are interpreted as furnishing direct empirical evidence of consciousness in animal subjects. Although evolutionary and neurological continuity give us good reason to think that some other animals are conscious, it is notoriously difficult to test that hypothesis directly, or to tell how far consciousness extends into the animal kingdom. We aim to show that conclusions about animal consciousness can be drawn from experiments on metacognition. Our focus is metamemory: an individual's ability to keep track of whether she accurately remembers a stimulus. We take metamemory as an illustrative case. It is not the only way in which conclusions about animal consciousness can be based on experimental observations, but by working through various methodological and philosophical objections in detail in this one case, we hope to demonstrate the merits of the broader methodology for investigating consciousness that we propose.

Since our focus is on evidence for consciousness, we do not aim to review the large comparative literature on metamemory, let alone metacognition in general. Of the many sorts of metamemory that have been studied, we are interested in the type of metamemory that can play an additional role, forming a plausible basis for inferences about consciousness. Which type of metamemory is indeed good evidence for consciousness—which type will do the trick?

To do the trick, the ability must be characterised in non-consciousness-involving terms *C*, in a way that makes it plausible that a subject's meeting condition *C* in relation to a perceptual stimulus is good evidence that they consciously remember it. Testing for condition *C* will then be one empirically-tractable way to probe whether other animals are conscious. Even those who reject higher order thought as necessary for consciousness should accept that some type of meta-representation can be evidence of consciousness. This paper addresses the question: what variety of meta-representation is suited to playing that evidential role?

The target of our investigation is phenomenal consciousness—the "what it's like"-ness of our mental lives. When we reflect on consciousness from the first person perspective, it can seem as if explaining and investigating it further is intractable. The logic of the approach taken here is to focus on what conscious experience does for us—to look for ways of characterising its functional profile. For example, there is evidence that there are two different ways of forming an association between a tone and a puff of air to the eye so that the tone comes to cause an eye blink: 'delay conditioning' and 'trace conditioning'. It seems that trace conditioning requires consciousness, whereas delay conditioning does not. In delay conditioning a puff of air to the eye is administered during the occurrence of a tone (after the start of the tone, hence ‘delay’ conditioning). Delay conditioning dissociates from awareness of the contingency between tone and air puff (Perruchet 1985). By contrast, it seems that trace conditioning—where the air puff occurs shortly after the tone has stopped—correlates with subjects’ conscious awareness of the contingency (Clark et al. 2001; Clark and Squire 1998; Perruchet et al. 2006). If it were established that only trace conditioning requires consciousness, then the presence or absence of trace conditioning, and of the mechanisms which underlie it, could be used as evidence as to whether other animals are conscious. When following this method, it is important that the mechanism of trace conditioning in humans be characterised in detail: its functional profile, the brain mechanisms involved, modes of intervening on or interfering with those mechanisms, etc. Such a detailed characterisation C_1 of the mechanism of trace conditioning is much richer than the bare observation that trace conditioning appears to correlate with verbal report of the contingency. The process of testing whether C_1 is present in other animals is correlatively more empirically tractable (and falsifiable).

We call a task ‘consciousness-involving’ if humans’ performance of the task, or their performance of the task in a particular way, correlates with their being conscious of the task-relevant parameters, as indexed by subjects’ introspective and environmental reports. We can study a range of consciousness-involving tasks (Jack and Shallice 2001). Given thorough investigation, the mechanism deployed in each consciousness-involving task can be characterised in detail: C_1, C_2, …, C_n. Each such characterisation is then susceptible to independent investigation in animals, without relying on verbal report, to see which other animals have the C_i mechanism. The purpose of this paper is to arrive at such detailed characterisation in the case of metamemory.

We should distinguish three types of potentially conscious state. First, there is the online visual perception of a stimulus. Second, there is visual recall of a recent past stimulus—the kind of state you are in when you shut your eyes and visualise the scene you have just been looking at. Third, there is metamemory: some kind of representation of your own visual memory. We ask whether states of the second kind are conscious, and focus on whether states of the third kind are good evidence of such consciousness. Information about an immediately past stimulus may be held online without being conscious. The claim we are considering is that there is a type of metamemory that correlates with the perceptual memory trace being conscious. Information about an immediately past perceptual stimulus is clearly a form of memory in the broadest sense. Where it falls in relation to the standard taxonomy of memory partly depends upon whether it is conscious. If so, it would be explicit rather than implicit. It would also be declarative rather than procedural, although that is a distinction that is usually applied to long term memory, rather than the short term memory involved in keeping information about an immediately past perceptual trace online. It is also episodic in character. Indeed, debates about whether long term episodic memory is evidence for consciousness in other species (e.g., Tulving 2005) have a similar structure to the issues considered here: although experiments may uncover behaviour that depends upon information about the time and place of some particular event in the animal's past experience, the further question arises about the circumstances in which the use of such episodic information provides evidence of consciousness. The process we engage in here of examining the metamemory literature with an eye to evidence about consciousness needs to be repeated for other tasks, like those involving episodic memory, that might also allow for direct empirical testing of consciousness in animals.

We will not discuss the neural mechanisms of human consciousness-involving metamemory, although they may be an important part of the story, but will instead aim at a broadly functional characterisation that can be carried across from humans to other animals. We call this condition *C*. Our aim is to formulate the condition *C* appropriate to consciousness-involving metamemory: what species of metamemory goes with a human subject's having a conscious perceptual memory of a stimulus? Finding that other animals do indeed satisfy such a detailed functional condition, in some circumstances, would then be good evidence that they were conscious in those circumstances; evidence that could be further reinforced by data about neural mechanisms (which there is not space to discuss here).

In formulating our condition *C*, we must walk a narrow ridge between tempting mistakes of opposite kinds. On the one hand we might formulate a condition which is in fact met by unconscious systems. That danger can be addressed by rigorous studies in people to ensure that the presence or absence of our proposed condition *C* does in fact correlate with the presence or absence of consciousness as measured by verbal report. However, while we concentrate on avoiding the abyss of the unconscious on one side, we may stray onto the comfortable slopes on the other where consciousness is a decidedly human-only phenomenon. That is, to be sure people only meet condition *C* when they are conscious we may formulate a condition that is too strong, which we know only humans can meet, effectively presupposing that animals cannot be conscious. For example, we might build verbal report or its equivalent into condition *C*. The discipline of formulating a condition *C* for which animals can be tested empirically should help to avoid settling on a condition whose connection with consciousness is exemplified only in humans. That is our objective. We aim to formulate a condition *C*, associated with metamemory tasks, for which non-human animals can be tested. The animal focus serves to ward us off the comfortable slopes of anthropocentrism.

It follows from the logic of our approach that finding behaviour in animals which is analogous to consciously-produced human behaviour has little forensic merit. Showing that an animal can solve a problem that a human would solve using metamemory casts little light on whether the animal is conscious. Experiments must test whether humans and animals deploy the same mechanisms, our focus here being on a functional characterisation of those mechanisms. In the next section, "Animal data", we give examples of work on metamemory that has moved towards this more stringent objective. In the following section, "Meta-Memory: high level meta-representation", we specify the type of metamemory which would be good evidence for consciousness and set out how it can be tested in animals.

Animal data

Cowey and Stoerig (1995), Stoerig et al. (2002)

From the extensive literature on metacognition in non-human animals (Smith et al. 2003), we select two experimental paradigms, each to illustrate a particular point. The first, which we discuss in this section, was deployed by Cowey and Stoerig in an elegant series of experiments on blindsight in monkeys (Cowey and Stoerig 1995, 1997; Stoerig et al. 2002). The second, discussed in the next section, is a memory discrimination procedure used by Hampton (2001) to provide evidence of meta-representation in a rhesus monkey. Both paradigms are based on the 'commentary key' method devised by Weiskrantz (1986, 1995).

Cowey and Stoerig studied monkeys with unilateral lesions of the primary visual cortex comparable to those which, in humans, give rise to blindsight—voluntary responding to visual stimuli in the absence of phenomenal consciousness. These lesioned animals were compared with intact controls on two successive tasks. In the first, 'localisation' task, the monkeys were rewarded with food for touching the visual target location, and the test stimuli were presented equally often in the right hemifield, where one would expect lesion-induced impairment, and in the left hemifield, where one would expect performance to be unaffected by the lesion. The results indicated that, at appropriate stimulus intensities, the lesioned animals could localise the stimuli presented to their right 'blind' field with almost 100% accuracy. The second, 'detection' task introduced the commentary key. In 50% of trials during initial training on this detection task a visual target was presented in the normal field and the monkey was rewarded if it touched the target location. The other 50% of trials were blanks, i.e., no target was presented, and the animal was rewarded if it touched a box stimulus that was constantly present on the computer screen. According to the logic of the commentary key method, touching this box constituted a report by the animal that it had not seen a visual stimulus in that trial. Once this discrimination had been mastered—once the animals were reliably touching the target on target trials and the box on blank trials—visual targets in the right 'blind' field began to be presented in 5% of trials. In these crucial probe trials reward was programmed for delivery whether the animal touched the probe or the box. The result was that the normal monkey consistently touched the probe, but the lesioned animals nearly always (92–98% of trials) touched the box. So, in combination, the two tasks showed that, when reward depends on it, monkeys with striate cortex lesions can localise visual stimuli in the 'blind' hemifield, but that when they have the option of getting reward without localisation, they act as they have learned to do when no stimulus was presented.

Cowey and Stoerig's findings show that, if monkeys are conscious, they exhibit blindsight in much the same way as human subjects. But that is to make the (plausible) assumption that some non-human animals can be conscious, not to test it. Cowey and Stoerig's studies do not demonstrate, or seek to demonstrate, that intact monkeys are perceptually conscious of the visual stimuli to which the respond. This is a perfectly reasonable assumption in the sense that it accords with most people's intuitions, and it is put to good scientific use in their research. It is used to test a 'continuity' hypothesis, the idea that striate cortex lesions have the same effects in humans and monkeys, against an alternative 'encephalization' hypothesis, which suggests that hominid evolution involved migration of visual function within the brain, and therefore that these lesions will have different effects in the two species. *If* one assumes that intact monkeys are conscious of the visual stimuli to which they respond, then Cowey and Stoerig's results support the blindsight hypothesis, with loss of consciousness following striate cortex lesions. However, if one questions this assumption, in the way that is necessary when the purpose of enquiry is to find out whether animals are conscious, its justification turns out to depend, not on careful functional analysis of visual perception, but on reasoning by analogy from one's own case (Heyes 2008). When I respond to visual stimuli I tend to be conscious of them, therefore when a monkey responds to similar stimuli under comparable conditions, I assume that he is also conscious of them. Below we argue that a more secure inference should be based on obtaining a detailed functional characterisation of how humans perform the task when they do so in a consciousness-involving way, and carrying that over as the basis of tests in other animals.

Our principal concern in this section has been to point out that Cowey and Stoerig's work, and other research that makes good scientific use of the assumption that nonhuman animals are conscious, does not furnish strong evidence that other animals are conscious.

Hampton (2001)

Hampton (2001, Experiment 3) used a memory discrimination task to produce good evidence for meta-representation in rhesus monkeys. Although he disclaimed any attempt to be studying the subjective experiences of his animal subjects (p. 5359), claiming that the *experiences* associated with remembering cannot be studied in non-human animals (p. 5362), we argue that Hampton's method can form the basis of experiments that would furnish evidence about this deeper issue.

At the beginning of each trial in Hampton's procedure, the monkey was shown one of four pictures on a computer screen (a new set each day). After picture presentation, there was a delay of variable duration (12.5–200 s), in which the screen was blank. After the delay, the monkey was usually required to touch one of two flags on the screen. Touching the 'test flag' resulted in the monkey being presented with a display containing all four pictures. If he selected from this array the picture he had seen at the beginning of the trial, he received a preferred reward, a peanut. By touching the other 'escape flag' the monkey could avoid the test but be sure of a lesser reward, a pellet of ordinary primate food (Fig. 1).

The result of the experiment was that the frequency with which the monkey chose the escape flag over the test flag increased with the duration of the delay after the original picture was presented. This pattern is consistent with the use of meta-representation by the monkey of its perceptual memory. That is, because memories fade over time, one would expect the probability of choosing the escape key to increase with delay if the monkey's decision whether to press the test key or the escape key depended on the strength of an internal representation of the sample stimulus. However, this relationship between choice of the escape key and delay

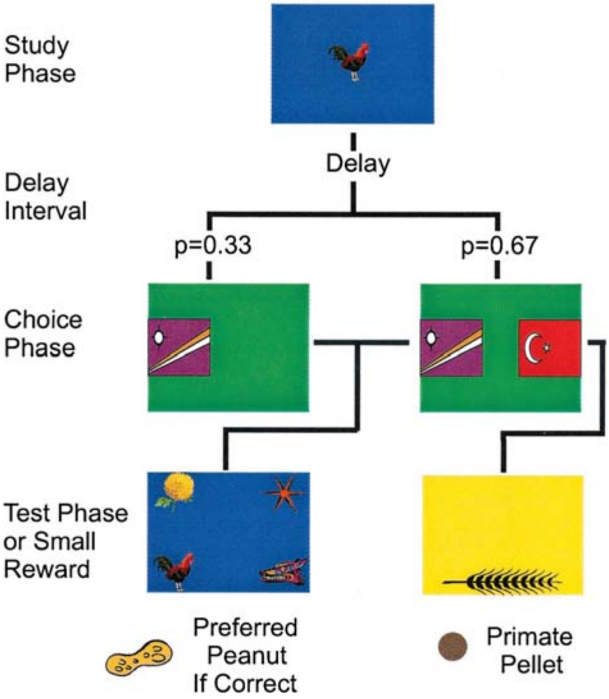

Fig. 1 The task structure in Hampton (2001)

duration is also consistent with the monkey having based his decision on the duration of the delay since the original image was presented. He may simply have learned that test taking tends to have a happy result when the trial started a short time ago, and an unhappy result when it started a long time ago. That would be to base the decision on a first-order representation of the delay rather than a meta-representation of the memory trace.

To test the first-order representational hypothesis against the meta-representational hypothesis, Hampton compared the accuracy of the monkey's performance on trials like those described above when he chose to take the test, with trials when he was forced to take the test. Approximately one-third of trials at each delay duration were forced trials. In these forced trials, only the test flag appeared at the choice stage. Hampton found that, when tests were forced, the proportion of trials in which the monkey correctly identified the sample picture declined as the duration of the delay increased, but when tests were chosen, accuracy remained high even at longer delays. If, in choice trials, the monkey's decision to press the test key or the escape key had been based solely on delay duration, one would have expected to see the same relationship between accuracy and delay duration in chosen and forced trials. Given that it declined with increasing delay in forced trials—presumably because the monkey was, in fact, forgetting the sample—then it should also have declined, at the same rate, in chosen trials because, by hypothesis, the monkey was using choice trials only to avoid taking the test at longer delays, and this could not be done in forced trials. The observed pattern—preserved accuracy at longer delays in chosen trials—implies that the monkey's use of the escape key enabled him, at *each* of the longer delays, to avoid taking the test when he was unlikely to pass it. The strength of the monkey's memory trace on any given trial will have been a powerful determinant of whether he was likely to pass the test. Therefore, having excluded the possibility that the monkey was basing his decisions solely on delay duration, the accuracy-preserving effect of his decisions makes it plausible that they were modulated by the strength of the monkey's own memory of the stimulus.

Foote and Crystal (2007) recently reported evidence of metamemory in rats. Although their procedure is similar to Hampton's in many respects, there were two crucial differences: the equivalent of the escape key (a hole into which the rat could poke its nose) was present, but not illuminated, on forced trials; and the rats made their discriminative responses at constant locations over trials. The first of these deviations from Hampton's procedure means that weaker performance in difficult forced tests than in difficult chosen tests could have been related to erroneous entry to the escape hole on forced trials. The second raises the possibility that each rat's choice of the test key versus the escape key was regulated by the position of its body relative to the response keys—an object-level, public motor cue. The potential for use of public motor cues—stimuli generated by the animals' own movements, rather than an internal representation—also makes the results of 'confidence judgement' experiments less compelling than Hampton's. In experiments of this kind (e.g., Kornell et al. 2007) commentary key responses are made immediately after the monkey has made its discriminative response. Therefore, commentary key responses could be controlled, not by an internal representation, but by some publicly observable feature (e.g., latency) of the preceding response.

It would be good to see a replication of Hampton (2001), in which the same result was obtained for more than one monkey, but since we don't think meta-representation is sufficient for consciousness, further experiments would be needed in any event to show that this is the type of metamemory that does the trick. We explain in section "Testing for Meta-Memory" below what kind of additional empirical work should be done to test for consciousness-involving metamemory.

Meta-Memory: high level meta-representation

Low level meta-representation is not sufficient for consciousness

We saw in the last subsection that Hampton (2001) offers evidence that a rhesus monkey can solve the memory discrimination task without relying on an external cue, instead using the strength of its own perceptual memory state as an internal cue for whether or not it is likely to succeed on a subsequent matching-to-sample task. We agree with Hampton that that is not yet evidence for consciousness in monkeys. However, we argue in this subsection that Hampton's result is indeed evidence for meta-representation.

As Reder (1996) has argued, there are important differences in the literature about what 'metacognitive' amounts to. Perner (1991) identifies meta-representational states in terms of their contents: they are representations of ‘representational relations’. Similarly, we take meta-level contents to be those which concern the thinker’s own representational states (e.g., *I am visually representing a red rose on the table*). Object-level states do not (e.g., *there is a red rose on the table*).

The results of Hampton’s experiment suggest that the monkey had an internal state, connected to its opt-out behaviour, which covaried with the strength of its recall of the visual stimulus, irrespective of which particular stimulus was being recalled. This accords with Hampton’s suggestion that the monkey might have an internal “flag” for the presence or absence of a memory (p. 5362). Why think the internal “flag” is a representation at all? Shea (2007) argues that meeting the following condition is sufficient to count as a species of internal representation: the animal acquires a new internal state R as a result of learning, the internal state R correlates with the state P of another system and leads to behaviour, and the behavioural output in response to R makes sense in the light of the property P with which it correlates (to state it informally). The monkey’s internal “flag” is an R that meets that sufficient condition. The property with which it correlates happens also to be internal: the presence or strength of a perceptual memory trace.

What is the content of this representation? Intuitively, the monkey makes a mistake if it opts for the test when it doesn’t remember the stimulus, and it also makes a mistake when it opts out of the test when it does have an accurate memory trace of the stimulus. These intuitive correctness conditions line up with a plausible account of the function of the representation. Its purpose is to keep track of memories. Without offering a full-scale theory of content, these considerations still suggest that the most plausible content for this internal state is meta-representational—something like *I have a memory of a visual stimulus*.[1]

We have been working with a sufficient condition for being a representation that is relatively easy to meet. There is no reason to think of the representation as ‘cognitive’ in the way that term is typically used in the metacognition literature. Relatively simple systems like those found in computers, subpersonal brain processing and animal signalling contain states with correctness conditions or satisfaction conditions. They would not count as ‘cognitive’. None displays the kind of psychological sophistication of human beliefs and desires, say (although they may in other respects be more sophisticated). We call them low level representations. We use ‘low level’ versus ‘high level’ not as a value judgement, but to mark this particular kind of variation in psychological sophistication of the representational states. That distinction is orthogonal to the distinction between object-level and meta-level contents. High level representations like beliefs can have both object-level contents (*there was a red rose*) and meta-level contents (*I can remember a red rose*). Low level representations may also have meta-level contents (e.g., when a computer keeps track of its own memory registers).

As we use the term, low level representations are non-conceptual, need not be in the space of reasons, may be coarse-grained, need not be at the personal level, and do not divide into a range of different mental attitudes such as believing, desiring,

[1] Although we use a structured (linguistic) representation to convey the content, we are not suggesting that the monkey’s representational state has constituent structure or conceptual content.

imagining and intending. Carruthers (2008) offers an account of animal performance in many metacognition experiments in terms of belief and desires with merely object-level contents. These explanations nevertheless presuppose a relatively high level of psychological sophistication because they attribute to animals states with the functional sophistication of beliefs and desires. The considerations we have offered suggest that Hampton's monkey may have a meta-variety of a relatively low level species of representation. Relatively little is needed for a new state driving rewarded behaviour to count as a meta-representation according to the sufficient condition mentioned above. Such meta-representations can arise in systems like current computers which are very unlikely to be conscious. Accordingly, we argue that this kind of low level meta-representation is not, on its own, good evidence for consciousness. Something must be added if it is to be turned into a plausible candidate for our condition *C*. In this subsection we investigate possible additions, to arrive at a characterisation of a metamemory mechanism which is plausibly consciousness-involving, which we call 'high level Meta-Memory', or just 'Meta-Memory'.

To turn it into an appropriate condition *C*, we need the mechanism for meta-representation to meet some further conditions: condition $C =$ low level meta-representation *plus* $X + Y + Z$. As we've said already, some of these conditions may concern neural structures and processes, for example particular brain areas (maybe only meta-representation in the prefrontal cortex is consciousness-involving) or processes (perhaps synchrony at the gamma-wave frequency is required). Since our aim is a functional characterisation of the consciousness-involving mechanism, we focus on additional requirements that can be characterised functionally (the X of $X + Y + Z$, as it were).

Philosophers have proposed various candidates for this additional factor X. Some argue that the meta-representation must have propositional structure (Rosenthal 2005). But to have thoughts with propositional structure requires a sophisticated capacity with powers akin to those of linguistic processing, which returns us to the problem of tying consciousness too tightly to something like language, and so ruling out animal consciousness at the start of the enquiry. Rosenthal does not claim that only creatures with language can have thoughts with propositional structure, but without the capacity for linguistic communication, it is very hard to gather evidence that an animal does have thoughts with propositional structure. In particular, the ability to categorise together a range of different stimuli, which is often the basis for studying 'concepts' in non-linguistic animals, is consistent with such generalisation behaviour being mediated by propositional structure or by non-conceptual representations. As a result, a test of consciousness that relies on finding representations with propositional structure would be hard to apply in non-linguistic animals. It is possible that the potential for consciousness does, in fact, depend on the capacity to give a verbal report, or on some important correlate of that ability—for example, possession of a language of thought, or the potential to code mental contents in propositionally-structured form. However, the case in favour of language-dependence and its cousins is not currently so strong that it justifies blank denial that animals are conscious. If we were to make this assumption, we would not only risk a major Type 2 error (concluding the phenomenon is absent when it is present), but also miss an opportunity to use the elucidation of conditions for the investigation of animal consciousness to clarify and extend theories of consciousness.

In searching for the additional factor *X*, other philosophers argue that only fine-grained contents are made conscious by meta-representation (Carruthers 2000). That proposal depends upon all perceptual experiences having such fineness of grain (cf. the experienced location of a touch on your back). But it does attempt to connect with the kinds of conscious experience that seem, from the first person perspective, to be involved in memory discrimination tasks, even if their fineness of grain is a contingent feature with respect to their being conscious. A third proposal is that Meta-Memory brings the perceptual memory into the space of reasons (McDowell 1994), making it available for the rational control of action (Baars 1988; Dehaene and Naccache 2001).[2] That, too, connects with an intuitive first-person perspective on memory discrimination tasks. When we use our own conscious recall of the perceptual stimulus to form an internal prediction of whether or not we will be able to match-to-sample, and then act on that prediction, it seems that information about the stimulus is thereby available to inform any kind of voluntary action (it is 'poised' to be acted on in any of a variety of ways). We are not suggesting that this would be the only way that human subjects could perform a Hampton-style memory discrimination task. But we argue that it *is* plausible that, when subjects succeed on the task by making use of their conscious perceptual memory of the stimulus, information about that perceptual memory is available to be consumed by any action system.

It is notoriously hard to spell-out this seeming availability. It has been argued that it is a distinctive functional property of human declarative memory (which is taken to be conscious) that subjects are able to discern the presence and absence of such memories (Tulving and Schacter 1990). Global availability is related, but goes further. It has at least two aspects. First, the way I keep track of whether I remember the stimulus is not proprietarily connected to a particular external cue, but is modulated in the same way by quality of the perceptual stimulus, delay since the stimulus offset, distraction, etc. Secondly, the representation of whether I remember can be deployed in the control of a range of different actions, rather than being dedicated to the service of only one project. In short, considering the philosophical positions and reflecting on conscious memory discrimination tasks from the first-person perspective brings us to the following characterisation of a potential additional factor *X*. In performing a memory discrimination task in a way that depends upon my consciously recalling the perceptual stimulus, I seem to have a representation, cued by my own memory trace of the stimulus, which is tokened in a variety of situations, and which is available for the control of a range of different actions and could be deployed to different ends were I given a different task. Such a representation would likely count as a meta-representation, following the discussion above, but it is a meta-representation which, in addition, can be tokened in a variety

[2] The idea of incorporation in the space of reasons derives from McDowell 1994. The condition we arrive at below is closer to the global availability for the rational control of action discussed by Baars, and by Dehaene and Naccache, which is less demanding than McDowell's notion.

of different situations and deployed to control a range of different actions. That requirement, when added to meta-representation, turns it into a plausible candidate for a functional characterisation of a consciousness-involving mechanism.

In sum, our functional characterisation, susceptible to empirical investigation in other animals, of the mechanism deployed in Meta-Memory (i.e., high level consciousness-involving metamemory) is as follows.

> Condition *C*
> The subject[3] represents that she has a memory of a perceptual stimulus, where that meta-representation can be tokened in a variety of different situations and can be deployed to control a range of different actions.[4]

We have generated condition *C* by introspective reflection on our own case. That is only a weak source of evidential support. It is enough, however, to be a plausible basis for generating a hypothesis for empirical test. It is a substantial empirical issue whether condition *C* does in fact correlate in humans with conscious recall of a perceptual stimulus as measured by subjects' verbal reports. That is, do human subjects meet condition *C* only when they report being conscious, and are they ever conscious without meeting condition *C*? The argument above makes it plausible, but does not prove, that condition *C* may correlate with other measures of consciousness—which is enough to motivate a proper empirical investigation.

Testing for Meta-Memory

Our condition *C* is susceptible to empirical test. The question to ask is whether performance on one type of memory discrimination task transfers readily to other memory discrimination tasks; that is, to use a triangulation approach (Campbell 1954; Heyes 1998). Two categories of transfer test must be combined to show that an animal meets condition *C* in a Hampton-type memory discrimination task. The first category of experiments demonstrate decoupling of the metamemory ability from any particular perceptual cue. For example, does an ability to solve Hampton's task where memory is degraded by a delay between initial stimulus presentation and the matching task transfer to trials where the accuracy of memory depends instead on variations in the duration or the intensity of initial stimulus presentation? The second category of transfer task looks for output generalisation: the ability to make use of the metamemory in a range of different tasks. For example, we might ask whether the type of representation that regulates opt-out behaviour in Hampton's memory test (matching-to-sample) could also be used to guide behaviour in a

[3] We use 'subject' to refer to the organism or system which encounters the perceptual stimulus, has a memory of it, and tokens a representation of that memory. We do not presuppose that being a subject in this sense involves a sense of self.

[4] We are deliberately vague about the modal claim 'can be deployed', since the aim is to match the intuitive ease or difficulty with which information about conscious representations can be deployed in the rational control of a range of different actions. For present purposes, we do not need to complete the separate project of making that notion more precise. The rough idea is that the meta-representation could be used for new projects simply by the animal changing its preferences or by it moving to an environment with a different reward structure, without having to undergo further learning in the domain of keeping track of perceptual recall (i.e., without having to undergo further meta-representational development).

different memory test (non-matching-to-sample). It seems very plausible that these additional tests would be satisfied. If Hampton is right that the monkey's opt-out behaviour is driven by an internal "flag" tied to the perceptual memory trace, then all ways of degrading that memory trace (delay, stimulus duration, stimulus intensity) would have the same effects. But that still needs to be tested. And it remains an open possibility that the animals perform the experiment in an informationally-encapsulated way: their training may have allowed them to keep track of the way the memory trace varies in the given experimental set up, but without being able to carry that over to situations where the memory trace varies in other ways, or where the information about the memory trace has to be used for different actions.

Obviously, these transfer experiments would be far from trivial. It would take a huge amount of work to design and implement effective experiments testing for the presence or absence of a mechanism meeting condition *C* in even just one other species. But there is no difficulty in principle with carrying out such investigations. Our condition *C* is both plausibly consciousness-involving in humans and yet susceptible to empirical test in other animals. In the remainder of this section we outline in a little more detail some potential experimental paradigms.

Transfer across perceptual cues

Testing for transfer across situations would be relatively straightforward, and has been discussed previously in the experimental literature on metacognition in animals (e.g., Inman and Shettleworth 1999). For example, monkeys would first be trained on Hampton's task, in which the strength of the animal's memory for the initial stimulus is manipulated by varying retention interval, i.e., the delay between presentation of the initial stimulus and the point at which the monkey has the choice of touching the test flag or the escape flag. Then, once the monkeys were responding in a way that suggests high level Meta-Memory—choosing to escape more often on long than short delay trials—occasional probe trials would be introduced. In these probe trials, the retention interval would be fixed and of relatively short duration, but the duration of the initial stimulus would vary. Sometimes it would be very brief, making the stimulus hard to encode and therefore to remember, and on other probe trials the initial stimulus would be on the screen for a longer period, making it easy to encode and remember. If a monkey selected the test flag in a probe trial, he would proceed to the usual, four-choice matching-to-sample test, but the test outcome would not be contingent on his response; he would be rewarded (or not rewarded) regardless of the image he selected. Therefore, and crucially, the monkeys would not have the opportunity to learn across probe trials that initial stimulus duration predicts test outcomes. Under these conditions, if Meta-Memory is indeed driving the monkeys' opt-out behaviour in the main task, then one would expect them to opt-out more often in probe trials with short than with long stimulus durations. This would be expected because, according to the Meta-Memory hypothesis, opt-out behaviour in the main task depends on the strength of a memory trace, not on the duration of the retention interval per se, and therefore the animal's tendency to take the test when the memory is strong but not when it is weak should persist when memory strength varies with stimulus duration rather than retention interval.

As with any single experiment, the outcome of this experiment would not be conclusive. Non-contingent reward on probe trials would ensure that the animals could not learn in the course of the experiment that stimulus duration predicts test outcomes, but it is not impossible that the animals would have learned this from their day-to-day experience before the experiment began. Transfer in a variety of different types of probe test, in which memory strength was manipulated not only by stimulus duration but also by, for example, the presentation of distractors before or after stimulus presentation, would strengthen the case for Meta-Memory. However, to rule out the possibility that success on probe trials was due to pre-experimental learning about relationships between perceptual cues and test outcomes, it would be necessary to use a novel, and possibly invasive, manipulation. If monkeys showed transfer in probe trials where memory strength was manipulated by direct neurochemical or neuro-electrophysiological means, and if one included appropriate sham controls, then we could be confident that pre-experimental learning was not responsible.

Transfer across actions/outputs

In Hampton's experiment, monkeys touched a flag of one colour to take the four-choice memory test and a flag of different colour to escape the test. To assess transfer across action types, initial training would be followed by the introduction of probe trials with different response requirements. For example, the monkeys might be required to pull one of two levers, rather than to touch one of two flags, to make their choice, or the matching-to-sample test might be changed to a non-matching-to-sample test. In the latter case, two images would appear on the screen, the initial stimulus and an alternative, and the monkey would be rewarded only if he touched the alternative image. Naturally it would take a while for the monkeys to learn the new contingencies—that pulling the left lever activates the test, or that non-matching performance is required in the two-choice test—but if their performance on the main task depends on Meta-Memory, then eventually they should show the same tendencies in probe trials as in trials on the main task, i.e., to opt-out more often when the retention interval was long, and to show greater accuracy at longer intervals in choice trials than in forced trials.[5]

The payoff

Amongst many tasks that may be consciousness-involving, we have examined metamemory. Hampton (2001) shows that monkeys can predict whether they are themselves likely to succeed at a visual matching-to-sample task. Although not

[5] Even if monkeys passed all of these transfer tests, it could be argued that their memory state is merely 'driving' their choice behaviour; that it plays an important causal role in generating their behaviour, but not by virtue of being understood by the animal *as* a memory. If consciousness of a memory were thought to require understanding the internal state as a memory, then more demanding empirical tests would be required, like those developed in the literature on theory of mind in nonhuman animals.

conclusive, his results suggest that one of his monkeys used a meta-representation of its own perceptual recall, rather than any external cue, to perform this task. However, meta-representation is not, on its own, plausibly good evidence for consciousness. What more is needed? By setting ourselves the objective of finding a condition which is open to empirical confirmation and disconfirmation in animals, we have avoided anthropocentric answers, and thus conditions which may correlate only in humans with the presence and absence of consciousness. We labelled the result high level Meta-Memory: a subject's representation of her memory of a perceptual stimulus, where that meta-representation can be tokened in a variety of different situations and can be deployed to control a range of different actions. Meta-representations which meet that further condition are plausibly good evidence for consciousness. And first-person reflection on consciously-performed memory discrimination tasks suggests that our conscious recall of the perceptual stimulus does indeed meet this condition, although that prima facie case must be substantiated by further empirical investigation. Thus, the payoff from our investigation of animal consciousness is not just to show, in the face of methodological scepticism, that it is an empirically-tractable question. It has also led us to a sharper conception of the nature of consciousness itself, in humans and other animals, forcing us to specify in greater detail the functional profile of the mechanisms deployed by subjects when they rely on consciousness to solve a memory discrimination task.

Those who view higher order thought as necessary for consciousness can take our condition *C* as a candidate for upgrading meta-representation into a sufficient condition for consciousness. But taking condition *C* as partly constitutive of consciousness would join higher order thought theories in making consciousness a matter of having certain dispositions. Our claim is less controversial: that discovering Meta-Memory in animals would be good evidence that they are conscious. That is a substantial claim, of considerable interest whether or not higher order theories are right. But this story has a final twist. We have been assuming throughout that, to solve a memory discrimination task without using an external cue, a subject would have to use some additional internal state, over and above its perceptual memory. We argued that, if so, the new representation would likely be a meta-representation, rather than having object-level contents. Our concern was to see what needed to be added to meta-representation, to turn it into a plausibly consciousness-involving mechanism. However, once we've seen that additional factor *X*, we can ask whether meta-representation is a necessary part of the evidential condition *C*, or whether the factor *X* would, on its own, be good evidence for consciousness. Assessing the theoretical considerations in favour of that hypothesis and making suggestions for testing it empirically would be a paper in its own right. We restrict ourselves to observing that our factor *X* is similar to Dehaene's global workspace hypothesis (Dehaene and Naccache 2001), which is formulated as a necessary and sufficient condition for a mechanism to be conscious.

Acknowledgments For helpful discussion and comments, the authors would like to thank Tim Bayne, Alan Cowey, Martin Davies, David Shanks, Kim Sterelny and an anonymous referee for this journal. NS's research is supported by the OUP John Fell Research Fund, the James Martin twenty-first Century School, the Oxford Centre for Neuroethics and the Mary Somerville Junior Research Fellowship, Somerville College.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Baars B (1988) A cognitive theory of consciousness. Cambridge University Press, Cambridge

Campbell DT (1954) Operational delineation of "what is learned" via the transposition experiment. Psychol Rev 61(3):167–174

Carruthers P (2000) Phenomenal consciousness: a naturalistic theory. Cambridge University Press, Cambridge

Carruthers P (2008) Meta-cognition in animals: a skeptical look. Mind Lang 23(1):58–89

Clark RE, Squire LR (1998) Classical conditioning and brain systems: the role of awareness. Science 280:77–81

Clark RE, Manns JR, Squire LR (2001) Trace and delay eyeblink conditioning: contrasting phenomena of declarative and nondeclarative memory. Psychol Sci 12:304–308

Cowey A, Stoerig P (1995) Blindsight in monkeys. Nature 373(6511):247–249

Cowey A, Stoerig P (1997) Visual detection in monkeys with blindsight. Neuropsychologia 35(7):929–939

Dehaene S, Naccache L (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition 79(1–2):1–37

Foote AL, Crystal JD (2007) Metacognition in the rat. Curr Biol 17:551–555

Hampton RR (2001) Rhesus monkeys know when they remember. Proc Natl Acad Sci USA 98(9):5359–5362

Heyes CM (1998) Theory of mind in nonhuman primates. Behav Brain Sci 21(1):101–114 (discussion 115–148)

Heyes CM (2008) Beast machines? Questions of animal consciousness. In: Davies M, Weiskrantz L (eds) Frontiers of consciousness. OUP, Oxford

Inman A, Shettleworth SJ (1999) Detecting metamemory in nonverbal subjects: a test with pigeons. J Exp Psychol Anim Behav Process 25:389–395

Jack AI, Shallice T (2001) Introspective physicalism as an approach to the science of consciousness. Cognition 79(1–2):161–196

Kornell N, Son LK, Terrace HS (2007) Transfer of metacognitive skills and hint seeking in monkeys. Psychol Sci 18(1):64–71

McDowell JH (1994) Mind and world. Harvard University Press, Cambridge

Perner J (1991) Understanding the representational mind. MIT Press, Cambridge

Perruchet P (1985) A pitfall for the expectancy theory of human eyelid conditioning. Pavlov J Biol Sci 20:163–170

Perruchet P, Cleeremans A, Destrebecqz A (2006) Dissociating the effects of automatic activation and explicit expectancy on reaction times in a simple associative learning task. J Exp Psychol Learn Mem Cogn 32(5):955–965

Reder LM (1996) Different research programs on metacognition: are the boundaries imaginary? Learn Individ Differ 8(4):383–390

Rosenthal D (2005) Consciousness and mind. OUP, Oxford

Shea NJ (2007) Consumers need information: supplementing teleosemantics with an input condition. Philos Phenomenol Res 75(2):404–435

Smith DJ, Shields WE, Washburn DA (2003) The comparative psychology of uncertainty monitoring and metacognition. Behav Brain Sci 26:317–373

Stoerig P, Zontanou A, Cowey A (2002) Aware or unaware: assessment of cortical blindness in four men and a monkey. Cereb Cortex 12(6):565–574

Tulving E (2005) Episodic memory and autonoesis: uniquely human? In: Terrace HS, Metcalfe J (eds) The missing link in cognition. OUP, Oxford, pp 3–55

Tulving E, Schacter DL (1990) Priming and human memory systems. Science 247:301–306

Weiskrantz L (1986) Blindsight: a case study and implications. OUP-Clarendon Press, Oxford

Weiskrantz L (1995) The problem of animal consciousness in relation to neuropsychology. Behav Brain Res 71:171–175