Animal Behavior and Cognition
ABC 2019, 6(4):289-308
DOI: https://doi.org/10.26451/abc.06.04.09.2019

What Do We Need to Know to Know that Animals are Conscious of What They Know?

Gary Comstock

Department of Philosophy and Religious Studies, North Carolina State University

*Corresponding author (Email: gcomstock@ncsu.edu)

Citation – Comstock, G. (2019). What do we need to know to know that animals are conscious of what they know? *Animal Behavior and Cognition, 6(4),* 289-308. https://doi.org/10.26451/abc.06.04.09.2019

Abstract – In this paper I argue for the following six claims: 1) The problem is that some think metacognition and consciousness are dissociable. 2) The solution is not to revive associationist explanations; 3) …nor is the solution to identify metacognition with Carruthers' gatekeeping mechanism. 4) The solution is to define conscious metacognition; 5) … devise an empirical test for it in humans; and 6) … apply it to animals.

Keywords – Metacognition, Animals, Introspection, Consciousness

Are metacognizing animals *conscious* of what they know? The question strikes me as a non-starter because I think metacognition is a conscious process by definition. As an ethicist interested in the ways we treat nonhuman animals, I first learned of the possibility that rats know what they know from Foote and Crystal (2007). According to them and many others, metacognition is thinking about thinking, or cognition about cognition.[1] To think about the contents of someone's cognitive states, much less *my* cognitive states, seems very hard for me to do if I am not conscious of what I am thinking about. In the latter case, I am thinking about me. How could I not be conscious of what I am doing?

In this paper, I argue for retaining the idea that metacognition is always a conscious state. But let me begin by distinguishing two kinds of cognition, first- and second- order. Suppose when I look at the following squiggles on a page, *Morgan's Canon,* I see only two capitalized words. I am not sure that I remember what they mean. In this first-order state, I am using my perceptual skills to direct attention to objects in my environment: black ink marks and what they represent. If I wish, I can go on to think about myself thinking about whether I remember the definition of Morgan's canon. If I do, I engage in metacognition, directing my attention to representations that are, or are not, in my memory. Metacognition involves more than first-order perception and comprehension. It involves a second-order "cognitive executive that supervises (i.e., oversees and facilitates) thought or problem solving" (Smith, 2005).

Can I use my cognitive executive to supervise—to monitor and control—my thoughts without being conscious of what I am thinking? I do not think it possible. How could I think about myself thinking without being conscious of who is doing the thinking? How can I cognize about my cognizing without cognizing about myself? In fact, if someone wants to argue that it is possible to metacognize without being conscious of what one cognizes, then they owe us an analysis that shows how that state might come about. Otherwise, the question that opens this article is moot, redundant. Of course animals

[1] I use "metacognition" to refer to cognition about my cognition and "mindreading" to refer to cognition about others' cognition.

are conscious of what they know if they are metacognizing! This assumption struck me as obvious when I first started reading the literature, and it is an assumption we should all accept if we do not already. Or so I shall argue.

The Problem is that Some Think Metacognition and Consciousness are Dissociable

The problem is that not everyone agrees that the analysis is correct. Objectors have argued that animals can metacognize without being conscious of what they know. These critics grant that I cannot turn my attention to my thoughts without being conscious of myself. However, they point out, I can turn my attention to my thoughts without *feeling* that I am doing so. Suppose the objectors are right. The next question is whether animals can do the same thing, metacognize without feeling that they are doing so? This is a legitimate question, and it is the one I want to address.

To get clearer about the question, let us review what we know. When a human metacognizer uses meta-representations consciously to reflect on the accuracy of her own representations, she employs ideas she has learned from others in order to categorize and report on her own judgments. We can measure her progress in learning, for example, which displays on a computer screens are "dense" and which are "sparse." Humans can be trained to press the "D" key in response to screens with a large number of pixels illuminated, and to press the "S" key in response to very lightly lit screens. When the subject subsequently faces an easy trial, she will respond quickly and correctly. She will respond more slowly and uncertainly to screens that are ambiguous between dense and sparse.

As the subjects learn the patterns the researchers make the tasks harder. Eventually the subject is given a third option, to bail-out of trials they do not want to answer. To select the third option is to refuse to use either of the task's primary discrimination options. When a human bails out we assume this means that they are uncertain. Using it means that they score a "miss" because they do not get an immediate reward in the event they guess correctly. But it also means that they avoid a penalty, say, a time-out period, in the event they guess wrongly. By choosing the escape option, subjects optimize rewards by skipping difficult trials.

We expect different people to have different levels of success at such tasks. While Sherry may be adept at scoring hits on a range of tests, Shirley may not be as skilled in learning subtle differences between screens. Making these kinds of judgments, in what are called type 1 tasks, need not involve anything metacognitive. The differences between Sherry and Shirley are at the level of perceptual sensitivity and visual discrimination. One is better than the other at this particular task.

We expect to find similar differences among animals in type 1 tasks. As readers of this volume will likely know, individual animals of different species are able to learn to discriminate dense and sparse screens. However, like Sherry and Shirley, animals are different and learn the skills at varying rates and with varying degrees of success. Our interest here, however, is not in these first-order processes in humans or animals. To explore a human's or animal's metacognitive potential, we must ask each participant to assess their level of confidence in their answers. In so-called type 2 tasks, subjects assess the state of their knowledge and deliver a verdict as to whether they have high, low, or no confidence in their guesses. Here, differences in personality and cognitive style may turn the tables. In type 2 tasks, Shirley may be a better judge of the accuracy of her responses than is Sherry. This may be true even though Sherry is better at type 1 tasks and gets more correct answers. In this example, Shirley is the better metacognizer.

There is a further confounding variable. In addition to being more sensitive in the area of self-knowledge than Sherry, Shirley may also be more (or less) disposed to believe she is usually right about things than Sherry. When someone generally has high confidence in their judgments, this response bias may operate independently of how well they assess what they know. To keep these two factors—accuracy about one's state of knowledge and confidence about one's accuracy—separate, experimenters use signal detection theory to measure the difference between type 2 sensitivity and type 2 response bias (Maniscalco & Lau, 2012).

Turning from human to animal subjects, we must deal with the fact that animals cannot verbally report on their confidence levels. Animals cannot tell us when they think they are metacognizing. Consequently, experimenters must rely entirely on behavioral outputs to determine the temporal parameters—the "when"—of the target mental activity. To address the problem, researchers give animals the opportunity to wager on their answers. In nonverbal displays of confidence (or lack thereof) in what they know, many animals will double down on their answers to the easy questions, trying to maximize their rewards and minimize their time-outs, just as human subjects do. When reflecting on their answers to difficult problems, many animals learn to opt out, again reflecting the behaviors of humans. When animals display so-called uncertainty behaviors—when they chose either not to bet on themselves or to foreswear answering altogether—they seem to be metacognizing. For that is what we assume the human subjects are doing when they behave in similar fashion.

Positive verdicts that animals engage in metacognition are not confined to one field. Psychologists endorse the conclusion (e.g., Couchman, Coutinho, Beran, & Smith, 2010). So do animal behaviorists (Rosati & Santos, 2016) and philosophers (DeGrazia, 2009; Gennaro, 2009; Proust, 2009, 2010). The case for concluding from these experiments that dolphins and monkeys, for example, are metacognizers with "I-thoughts" is an argument from analogy (Gennaro, 2009). We presume that humans are metacognizing when we opt out. By parity of reasoning, therefore, when an animal behaves in a manner closely mimicking the human, we conclude that the animal is metacognizing, too. The idea is that both nonhumans and humans use low-level perceptual cues to answer easy problems because both are directly stimulated by unambiguous environmental prompts. However, nonhumans, like humans, shift to high-level "meta" processes to answer the harder problems because the questions are difficult, there are no direct stimuli, and so the subject must be surveying the contents of their cognitive repertoire. The animals, like the humans, are therefore conscious of their thoughts.

According to the argument from analogy, animals exhibiting uncertainty behaviors must be conscious of what they know–or, more precisely, what they do not know—just because humans exhibiting the same uncertainty behaviors must be conscious of what they know and do not know. So, now that we have the right question before us, we can ask the next one: What does it mean to be conscious of what one knows and does not know?

The answer to the question, "What is consciousness?" is unknown. Many answers are proposed and every one has its problems. In the current state of impasse, one approach has served as a starting point. To be in a conscious state is for there to be *something that it is like* to be in that state (Nagel, 1974). To have experiences is to do more than perceive. Seeing, hearing, and tasting are perceptual states, but consciousness is more. It is to combine what one sees with what one hears, to have *qualia* or subjective mental states in which perceptions from various modalities are combined in a single point of view. In accord with this beginning point, we may ask, "What does it feel like for a human to metacognize?" As I will soon explain, I do not think we should rest content with this question, but it is a good enough place to start. (We might well ask, where else could we start?)

A plausible answer is that metacognition does not always feel like one thing; it feels like different things on different occasions. If I am engaged in religious self-reflection, it may feel calm, meditative, perhaps even trance-like. If I am trying to remember the name of a college buddy, I may be caught in that maddening "tip-of-the-tongue" experience. If I am trying to escape intrusive thoughts, I may feel suicidal. If I am trying to decide whether a screen is dense or sparse, I may feel unsettled, mildly nervous. Metacognition can feel pleasant or unpleasant, have positive or negative valence, feel like confidence or puzzlement, and various combinations of these feelings. Metacognition occurs in a variety of circumstances with a variety of causes and profiles. The feelings I have just listed do not come close to exhausting the feelings that accompany my thinking about my thoughts.

That is my first point. Metacognition involves a range of feelings. And here is my second point. Metacognition involves some feeling or other. If you do not feel something while you are allegedly metacognizing, then you are not metacognizing. For while we can be, and usually are, conscious without metacognizing, we cannot be metacognizing without being conscious. We do not, for example,

metacognize in our sleep or under general anesthesia. Or so it must be if metacognition requires consciousness.

We can go further. Metacognition does not require simple consciousness. It involves a particularly sophisticated form, consciousness of *self*. As Smith and Washburn (2005) noted,

> Metacognition demonstrates humans' awareness of the processes and limitations of mind. *It is taken to show their self-awareness because uncertainty is so personal (i.e., you know that you feel uncertain).* Metacognition is linked to declarative consciousness (i.e., to the aspects of mind that humans have awareness of and can talk about) because humans easily introspect mental states like uncertainty and express them (p. 19, emphasis added).

The passage uncovers an assumption of many of the early metacognition investigators, including one of the first (Gallup, 1982). All things metacognitive are by their very nature conscious. It is an assumption I share. Anything worthy of the name metacognition is conscious just because metacognition is the ability to survey the contents of one's memory (Do I know what *Morgan's canon* is?) and to use the results of the survey to guide one's behaviors (I think I'll google *Morgan's canon* and check myself) (Koriat, 2007; Nelson, 1996). Metacognizing subjects cannot be unconscious of what they are doing if they are monitoring and controlling the contents of *their* consciousness.

So here is where I begin. It has seemed to many, and still seems to me, that consciousness comes for free with metacognition (Baars, 2003; Baars & Franklin, 2003). If an animal is metacognizing, she is conscious of what she is thinking. But this is only the beginning. The lay of the land has shifted.

In more recent rounds of animal metacognition research, investigators have called the link between consciousness and metacognition into question (Koriat, 2000; Kornell, 2009; Shea, 2019; Vandekerckhove & Panksepp, 2009). The argument that the two are dissociable holds that many unconscious processes underlie our conscious processes, that we are unable to direct our attention to these unconscious processes, and yet the unconscious processes are part of the conscious processes. Consider, for example, working memory. While we lack access to the mechanisms that support what goes in and what comes out of working memory, those mechanisms indicate conscious working memory and are properly considered a part of it (Dutta, Shah, Silvanto, & Soto, 2014; Hassin, Bargh, Engell, & McCulloch, 2009). Or so goes the argument. The mechanisms of other cognitive processes are also unconscious: consider activities such as reading, doing arithmetic, and responding to quiz questions.

Many metacognition researchers have adopted this way of thinking. Common usage now refers to nonconscious mental processes as "metacognitive" just in case they support conscious metacognition. Many authors distinguish "implicit" from "explicit," or "nonverbal" from "verbal," metacognition (Beran, Brandl, Perner, & Proust, 2012; Dienes & Perner, 2002; Koriat, 2000, 2007; Proust, 2013; Shea, in press; Vandekerckhove & Panksepp, 2009). Accordingly, metacognition now refers to processes both above and below the level of subjective awareness.[2] Here we reach the heart of the matter that motivates me to write this essay. Given current usage, metacognizing animals may or may not be conscious of what they are thinking. Everything depends on what the researcher is counting as evidence of metacognition and what the researcher takes consciousness to be. One can no longer assume that if data show an animal is metacognizing then the animal is self-conscious. For better or worse, we must now distinguish claims for conscious metacognition from claims for unconscious metacognition.

So, what counts as evidence for conscious animal metacognition?

---

[2] Not everyone finds these developments auspicious. Speaking personally, it seems to me that inquiry into "implicit metacognition" would be more helpfully deemed inquiry into *cognition*. I am also puzzled by the phrase "supra-personal metacognition." Supposedly a group of us can participate in a "System 2" broadcast of representations, broadcasts that transcend any of our individual consciousnesses and that serve, without our knowledge, to coordinate "the sensorimotor systems of two or more agents…" (Shea et al., 2014). Couldn't we call that research into *communication* or *group action*? One might be forgiven for worrying that the term metacognition is being stretched beyond its usefulness.

The Solution is Not to Revive Associationist Explanations

In this section I pause to clear away some brush. There is no debate about whether animals engaged in so-called *implicit* metacognition are conscious of the representations constituting their thoughts. Since "implicit" representations are by definition unconscious, these animals are by definition not feeling or attending to these processes. Whether such processes are "metacognitive" in name only is a semantic question I will not pursue. (I suspect it is.) But, going forward, I set aside any and all experiments claiming to show (only) implicit animal metacognition. They are irrelevant to my question. I focus instead on data that seems to show conscious metacognition in animals.

At this point it will be useful to break the history of these experiments into an earlier and a later round. I will argue that associationist explanations work for the data in the first round of experiments. But they do not explain the data from the second round.

In the earliest round of experiments (e.g., Smith, Shields, & Washburn, 2003; Smith & Washburn, 2005) animals underwent extensive training. They were given clear, individual stimuli in direct discrimination trials and correct responses were rewarded immediately with highly desirable consequences. After many weeks of preparation, the animals were presented with the opportunity to opt-out. In these experiments, animals did not need to reflect on their own state of knowledge before making their choices. They could rely on a habituated task-specific rule (e.g., "if my arm does not move swiftly after the stimulus to hit the dense or sparse key, then opt out") to evade an unwelcome outcome (e.g., "no food reward"). So, interpreters of the data from the first round of experiments could justifiably claim that animals only needed low-level associative strategies to perform in the way they performed (Smith, Beran, Couchman, & Coutinho, 2008). As critics have observed, metacognition is not the only available explanation of an animal's use of the opt-out key in these experiments. The animals' behavior can be explained in associationist terms (Church, Jackson, Beran, & Smith, 2019; Hampton, 2009; Jozefowiez, Staddon, & Cerutti 2009; Le Pelley, 2012). They point out that during the training period, the animal learns that her use of the uncertainty key can help her avoid aversive consequences. So, rather than using the key after she has metacognitively assessed her memory bank and determined that she is not sure of the answer, the animal, critics argue, is simply reacting to perceptual, behavioral, or heuristic stimuli. According to the critics, the animal's behaviors are another form of associative-reactive responding residing in associative-cue environments (e.g., Basile, Schroeder, Brown, Templer, & Hampton, 2015; Beran, Smith, & Perdue, 2013; Ferrigno, Kornell, & Cantlon, 2017; Smith, Church, Beran, & Washburn, 2018; Smith, Coutinho, Church, & Beran, 2013; Church et al., 2019).

Here is one way to construe the criticism. The experimental animal has three different external cues to which she can respond: the perception of a screen that is clearly sparse, or the perception of a screen that is clearly dense, or the perception of a screen that is neither sparse nor dense. When she sees a screen that approaches the category boundary between sparse and dense she also perceives *her body* engaging in what the behaviorist Tolman called "catalyzing behaviors," or "lookings or runnings back and forth" (Tolman, 1938). When the animal sees a clearly sparse (or dense) stimulus she immediately responds by pressing the sparse (or dense) key. But when she sees an intermediate, "ambiguous," screen she responds by freezing. She does not act at all. She defers, lollygags, procrastinates. This lagging behavior is part of the third stimulus, a stimulus visible to the researchers and visible to the animal herself.[3] Seeing her body's uncertain movements, the animal eventually responds by pressing the uncertain key. If this interpretation is correct, then the animal is not reading her *mental* state; she is reading her *bodily* state.

If an animal's apparently metacognitive performance can be explained as a learned association of a perceptual prompt with an immediate concrete reward, then allegedly metacognitive behavior may in fact be little more than mere responsiveness (Church et al., 2019). In sum, associationist explanations of

[3] I happened upon this fruitful way of putting the point in the literature somewhere but cannot now find the author.

the data found in early metacognition experiments make it unnecessary to attribute to the animal conscious introspection of her mind (Hampton, 2009). While I accept associationist explanations of the data in the first round of experiments, I do not accept them for the data in later experiments. In later, more complex, trials, animals acted in nuanced ways that associationist explanations do not explain (Smith, Couchman, & Beran, 2014).

In later experiments, new protocols made it impossible for animals to interpret their bodily responses as prompts to stimuli. Researchers devised protocols to rule out the possibility that the animals were choosing the bail-out option simply in virtue of its reward properties. For example, in one "block design" trial, animals did not learn which of their individual responses paired with which questions (Beran, Smith, Redford, & Washburn, 2006; Smith, Beran, Redford, & Washburn, 2006; Smith, Redford, Beran, & Washburn, 2010). Subjects were prevented from learning whether their answers to individual questions were right. Instead, they learned only whether they were on-balance correct in responding to a set of, say, four questions or, say, a variety of different kinds of tests. In these trials, animals did not receive immediate or high value rewards. In some tests, animals worked under heavy cognitive loads and had to rely on flexible, domain-general capacities rather than fixed action patterns and domain-specific rules. Since animals could no longer learn optimal responses by making associations between stimuli and perceptual cues, they could no longer rely on their reinforcement histories to identify aversive stimuli. Under such conditions, associationist paradigms are insufficient to explain the data (Couchman, Coutinho, Beran, & Smith, 2009).

The second round of experiments indicates progress in the attempt to identify reliable behavioral signals of metacognition. Consequently, we are closer now than we were two decades ago to being able to explain how an animal would act were it to be having mental experiences like the experiences we have when we metacognize. However, whether the resulting behaviors actually do indicate metacognition, either in animals or non-reporting humans, or whether they only tell us when an animal (or human) is acting *as if* they are metacognizing, is still to be established. The reason is that we do not yet have a firm grip on whether the behavioral signs of metacognition in humans indicate real metacognition. The case for other animals probably stands or falls on how this prior question is eventually resolved. If we do not metacognize, it is unlikely that other species do. And we may not.

In the criticisms to follow, as I say, I focus only on the later experiments, those that offer more robust support for the idea that animals are conscious of what they know. Are the animals' behaviors in these experiments to be explained by appealing to second-order representations that exist "inside" their minds? Or can the explanation be complete appealing only to first-order representations of things—including their own bodies—that exist "outside" their minds?

Nor is the Solution to Identify Metacognition with Carruthers' Gatekeeping Mechanism

In a series of articles, Peter Carruthers has argued that first-order, or world-directed, explanations are available for the data accumulated from the experiments to date (Carruthers, 2008, 2009a, 2011, 2014). His model makes three "folk psychological" assumptions.[4] The first assumption is that humans and animals have beliefs and desires. The second assumption is that beliefs and desires exist on a scale from weaker to stronger. The third assumption is that first-order beliefs and desires are not metacognitive (Carruthers, 2008, 2009a, 2009b).

With these assumptions in place, explanations need not be behaviorist because they can make liberal use of folk psychological concepts, concepts such as belief and desire. These concepts are not

[4] Folk psychology is a term of art in philosophy, referring to the semantics of the theoretical terms we ordinarily use to describe our mental states. On a widely accepted view, the meaning of common words such as "belief" and "desire" depends on the function of these words in a theory of cognition, and their theoretical function depends in turn on their causal role (see, for example, (Lewis, 1966, 1970; Ravenscroft, 2019)

available to behaviorists. So, on Carruthers' account, animals are not unfeeling automata responding robotically to stimuli. Rather, they have genuine desires for rewards, real beliefs about what they see, and determinate abilities to pause to gather additional information about what they are seeing (Carruthers, 2013). On Carruthers' view, animals want to obtain rewards and have stronger and weaker beliefs about how to do it. He postulates that animals also possess what he calls a gatekeeper mechanism. This mechanism equips animals to act on their strongest beliefs, to refrain from acting on their weakest beliefs, and to freeze the animal from acting should two beliefs of equal weight conflict with each other.

Carruthers' picture looks like this. When a trained animal is presented with a screen that is clearly dense, she draws on her past experience with similar screens to form a strong belief that if she presses the D key she will get a reward. Carruthers represents the situation using numbers to indicate first-order representational states along with subscripts to indicate the strength of the belief, weak ($_w$) or strong ($_s$).

(1) BELIEF_s [if the pattern is dense and D is pressed, then food results].

In this case, the animal sees that the pattern is dense and presses D. However, the situation is different when a more difficult screen appears, the kind of screen supposed to elicit metacognition. Now the animal has only a weak belief that the screen is dense, and this belief is matched in strength by a countervailing weak belief that the screen is sparse. Since the two beliefs are both weak, and since the animal does not want to make an error that will lead to a time-out, the situation is as follows.

(2) BELIEF_w [the pattern is dense].
(3) DESIRE_s [food].
(4) BELIEF_s [if the pattern is sparse and D is pressed, then a time out results].
(5) BELIEF_w [the pattern is sparse].
(6) DESIRE_s [no time out].
(7) DESIRE_w [press D].
(8) DESIRE_w [don't press D].

Since (7) and (8) contradict each other, the animal freezes, not being able to act on either desire. Note that, on this account, the animal's uncertainty is not caused by her consciously taking inventory of what she knows. Nor is it caused by her pausing to consider whether she knows the correct answer or by any other kind of self-assessment. Her behavior is determined directly by her not having a distinct perception of what she is seeing. She is seeing one object in two different ways. She is forming a weak belief that the pattern is dense (and, correspondingly, wanting to press D) while at the same time forming a weak belief that the pattern is sparse (and wanting to press S). Since pressing D means she cannot press S, and since pressing S means she cannot press D, and since she does not strongly believe she should press either key, she presses neither key. She is tied up, as it were, not by her second-order judgment that she does not know which key to press. She is tied up by the fact that she has two conflicting perceptions and a psychology that requires unifying them (Comstock & Bauer, 2018).

According to Carruthers' model, the animal's inability to act in the face of a difficult screen is due to the intervention of a blind referee, a switch Carruthers calls a gatekeeping mechanism. This is a domain specific module that has only one function, to determine which beliefs are eligible for conscious status. The mechanism adjudicates between beliefs, allowing stronger beliefs to be broadcast to other modules. As long as one belief is stronger than another, it will be admitted to the space of consciousness. However, when two beliefs are of equal strength, as in the example above, then neither belief will be allowed to influence action.

The gatekeeper works only on first-order representational states. It is not a part of the metacognitive process. Metacognition occurs only when the gatekeeper fails. Because animals must make decisions within a given timeframe, the gatekeeper only has a limited period within which to decide which beliefs to admit. If the gatekeeper receives an overload of inputs and is unable to adjudicate them before time is up, it may fail to forward any contents at all. Or, it may serially throw some string of its

contents, some of them potentially contradictory, into the space of consciousness where multiple higher-level operations will have to go to work to try to resolve them. This would be the earliest point in time at which metacognition could possibly begin. Metacognition cannot occur until the gatekeeper mechanism has either completed its work or broken down (Comstock & Bauer, 2018).

In response to Carruthers' model, Couchman et al. (2009) argue that the gatekeeper mechanism *just is* metacognition. Carruthers, they write,

> devises a secondary mental construct to explain why an animal uses the uncertainty response in too-close-to-call situations. He suggests that some species have a gate-keeping "mechanism . . . which when confronted with conflicting plans that are too close to one another in strength will refrain from acting on the one that happens to be strongest at that moment, and will initiate alternative information-gathering behavior instead" ([Carruthers, 2009a], p. 66). The gatekeeper mechanism operates on first-order cognition's outputs to assess their ability to produce a correct response. It meets the definition of a second-order controlled cognitive process. It produces a qualitative change in behavior and cognitive strategy (information seeking, uncertainty responses, etc.). It typifies the metacognitive utility that all theorists have envisioned (p. 142).

Let us first observe that Couchman et al. have not said that "a second-order controlled cognitive process" is one that produces "a qualitative change in behavior..." alone. We know that hormonal changes in one's body can produce qualitative changes in behavior, but hormonal changes are not controlled cognitive processes. What Couchman et al. have written is that metacognition produces changes in "in behavior *and cognitive strategy*." Is this what Carruthers' gatekeeper does? Hardly. The gatekeeper mechanism is part of the first-order process of determining whether one belief is stronger than another. That is all it does. It is an abstract postulate meant to explain which of two competing first-order beliefs may become conscious. It is, as Carruthers metaphorically characterizes it, a mechanical device.

Perhaps an inexact analogy with a mechanical device might help. The gatekeeper's functions, such as they are, are similar to the functions of the bi-metallic strips in a thermostat. When the temperature in the room becomes warmer, the strips expand and cause the device to register a change in room temperature. When looking at the device's display over time, an observer may become conscious of the information about temperature emerging from the thermostat, but the thermostat does not. Similarly with the gatekeeper. Observing the gatekeeper's work over time, we may become conscious of the information about belief emerging from the gatekeeper, but the gatekeeper does not.

Carruthers' critics mistake the gatekeeping mechanism for a second-order cognitive process. To the contrary, the mechanism is a hypothetical construct encapsulated in a first-order procedure that only compares the strengths of the beliefs that enter it. It determines which beliefs, if any, are eligible to enter consciousness. When a strong belief outweighs a weak belief, the strong belief is globally broadcast across the brain. That belief is now eligible to become conscious. But if that belief is itself to become the target of a thought, other mental operations must take place. Carruthers' gatekeeper supports cognition. It is not a part of metacognition.

The misunderstanding of Carruthers' model exemplifies the problem with which I opened this article. Metacognition now encompasses a wide variety of mental processes, many of which are not conscious. With such a loose definition of metacognition, it is not difficult to establish metacognition in other species. Carruthers' cognitive architecture is more demanding. It preserves the idea that metacognition requires consciousness. If first-order explanations of the data in the more sophisticated animal experiments are correct, then we need more carefully designed experiments if we are to establish conscious metacognition in animals. One may reply that the animals in the second round of experiments are engaged in "implicit" metacognition, but this reply will not satisfy those who think metacognition must be conscious.

What kind of experiment would show that an animal is conscious of what she knows and does not know? That depends on what we mean by consciousness. What is it for me to be conscious of some cognitive, emotional, or motivational state of mine? The answer requires scientific and philosophical

investigation. Scientifically, consciousness depends on the empirical facts about what is happening in the brain. Philosophically, it depends on how consciousness is defined.

The Solution is to Define Conscious Metacognition

Philosophers tend to divide into two groups when it comes to consciousness. *Substance dualists* believe that consciousness is a feature of an immaterial mind. Like Descartes, they hold that the mind is dissociable from and irreducible to the brain. Consciousness, therefore, is a basic, unanalyzable, concept. We can only speak about it in metaphors, explaining "what it is like" when, say, we wake up and become conscious of the alarm clock. For dualists, there is "something that it is like" to be me and to have my experiences. The subjective "feel" I have at the moment is equivalent to my self, and that is about all that can be said. This self is related to what is happening in my brain, but it is not reducible to it. Observers can discriminate objects that possess simple consciousness from objects that lack it by watching the object's motions. We can tell whether something can feel pain, for example, by presenting an aversive stimulus and assessing whether the response evidences avoidance and withdrawal, and whether these behaviors are modified by analgesics. Observers can discriminate subjects with complex consciousness from subjects that lack it by asking them questions, such as whether Sally will know to look under a bed for her shoes after someone has unbeknownst to her moved her shoes there. But Sally's being "conscious of" her shoes is not explicable in physical terms. This is the thought behind Thomas Nagel's (1974) claim that science will never be able to tell us what it is like to be a bat.

For our purposes, we can rule out substance dualism. If minds exist that float free of matter, the methods of the empirical sciences cannot reach, much less explain, them. Fortunately for us, another group of philosophers think consciousness is amenable to scientific exploration. Naturalistic approaches to consciousness tend to subdivide into five main theories: global workspace (Baars, 1988, 2005a, b); information integration (Tononi, 2004, 2008); higher-order thought (Carruthers, 2000; Lycan, 1996; Rosenthal, 2005); higher-order experience (Lycan, 1987, 1996); and first-order thought (Dretske, 1995; Goldman, 1993; Rowlands, 2001). A sixth view, reductive materialism, holds that the mind is either identical with the brain (Block, 1978; Oakley & Halligan, 2017; Smart, 1959) or an illusion altogether (Churchland, 1989; Churchland & Churchland, 1998; Dennett, 1991, 2005; Rosenberg, 2011).[5] On all of these accounts, the results of scientific inquiry are critical to understand consciousness.

A word about reductive materialism. On this view, consciousness does not exist apart from neural processes in the brain. In addition, it is explained away by the laws of physics. We do not currently have a complete explanation of consciousness, but that is because the physical sciences are not fully mature. When they are, they will provide detailed, causal explanations of all mental states, and at that point our folk psychological terms (such as "beliefs" and "desires") will be replaced by more precise neurological terms. A mature physics will produce testable hypotheses about what is happening in the brain when one, for example, feels anxious about their preparedness for class. Eventually, we will (probably) decide that our prior talk of "conscious states" was based on ill-informed illusions, illusions we would better off doing without. Reductive materialists think that replacing folk psychological explanations ("consciousness") with scientific explanations will be more accurate empirically and beneficial psychologically. It will, for example, allow us to understand and treat mental illness more effectively.

For non-reductive naturalists the opposition of folk psychology and materialism is unnecessary. If you hold that consciousness is best understood in folk dualistic terms as the mental states that make up one's self, and that all mental states supervene on physical states, then you may think that talk of mental states is compatible with a materialist metaphysics. *Compatibilists* are naturalists who believe that any change in one's mental state must be accompanied by a change in one's physical state. So, unlike dualists,

[5] Two additional theories are controversial. Roger Penrose's view that quantum mechanical effects account for consciousness (Penrose, 1994, 2016) has difficulty with the short time scale of neuronal firing (Tegmark, 2000). Panpsychist theories (Chalmers, 1996; Strawson, 2006), according to which human consciousness emerges from more basic forms of consciousness, have difficulties with the counterintuitive commitment of the theory that chemicals and pieces of DNA are conscious and with the problem of extracting one conscious subject out of many such conscious subjects (Goff, Seager, & Allen-Hermanson, 2017).

compatibilists do not believe in an immaterial mind untethered from the physical world. And, unlike materialists, compatibilists do not believe that we can jettison folk psychology or reduce mental states to physical states. Not yet anyway.

Whose account of the human mind/brain is correct? The question is unlikely to be decided soon. Due to limitations of space, I will not survey these theories (but see Van Gulick, 2018). Instead, I adopt the one that seems to have the most empirical support from neuroscience, global workspace theory, and its philosophical cousin, higher-order thought.

Global workspace theorists believe consciousness arises when the outputs of specialized brain sub-modules are made available as inputs to most, if not all, other modules. When information is "globally broadcast" across vast expanses of the brain, we become aware of it. With respect to vision, for example, information received initially in the brain's posterior V1 area is eventually sent forward to frontal, medial, and lateral cortical areas. Long distance transfers are supported by hierarchical neural connections making it possible for the raw data to be conveyed to distant areas, such as the thalamus and fronto-parietal cortex (Dehaene et al., 2001). These two-way streets enable us to send perceptual information to decision-making areas for processing. There it can inform signals relayed on to motor control areas to initiate movement. In this way, conscious subjects attend to what is perceived. We can manipulate and adapt inputs to suit our interests and use them as a basis for action (Baars, 2005b; Edelman, Baars, & Seth, 2005; Stoerig & Cowey, 1995). By allowing us to deliberate more slowly about the results of the fast and automatic limbic system, the global workspace allows us to become conscious of what we are doing and thinking (Herzog, Kammer, & Scharnowski, 2016; Baars, 2002).

Higher-order thought (HOT) theorists believe consciousness arises when lower-order percepts are processed by higher-order mechanisms. According to the theory Carruthers defends, Interpretive Sensory-Access (ISA), beliefs and desires are lower-order psychological states. Knowledge of them, however, is higher order. While we can know what we believe about the environment directly, we can know what we believe about our beliefs only indirectly. Unlike theories, such as Descartes' theory, that presume we are transparent to ourselves, ISA holds that we have no infallible access to our mental states. On Carruthers' view, our minds are opaque to us. We can only access our conscious propositional attitudes (for example, our judgments and decisions) in an indirect way. We must interpret our behaviors, sensory percepts, mental imagery, proprioceptive information, and internal speech, if we are to come to any self-knowledge at all.

For HOT theorists, consciousness arises only when sensory based inputs are broadcast globally because only then can they come under the control of executive structures in the frontal cortex. According to Carruthers' HOT version, we use a special faculty to interpret our intentions and judgments, the same faculty we use to decipher the cognitive states of other people. For Carruthers, our consciousness of sensory inputs can be transparent. Lower kinds of consciousness, such as discriminatory consciousness, allow us to recognize and categorize items in the world. Discriminatory consciousness, which animals share, enables one to know that one is seeing a canon and to be able to tell a canon from a razor. Another lower form, phenomenal consciousness, is the inner, subjective, qualitative state of awareness, Nagel's "what it's like" to be in some state (cf. Tye, 2003). These lower forms of consciousness do not require or employ second-order representations.

On the other hand, consciousness of the contents of our own minds requires second-order representations and the ability to process and interpret them. Only when fine-grained nonconceptual sensory contents are made available to the global broadcast system can HOT begin to occur. Only then can we begin to reason, make judgments, or think about what we know and do not know.

If ISA is the correct account of the way furniture is arranged in the human mind, then conscious metacognition cannot be directly introspective. It must await the indirect results of whatever the global broadcast system brings forth. If ISA is the correct account of consciousness, how should scientists interested in consciousness proceed? We must start somewhere. It is not clear that we have any other choice than to proceed with our current folk psychological categories. The reason is that reductive materialism requires a method for translating the findings of psychology to biology, and biology to chemistry, and chemistry to physics. We do not have this method yet. Currently our only starting point for

the explanandum of consciousness, the thing to be explained, is our folk psychological acquaintance with our own mind. Until objective, third-person accounts are developed to explain consciousness, our first-person, subjective language of belief and desire, pleasure and pain, love and shame, must suffice.

So the situation is this. Animal metacognition researchers only have recourse to the scientific methods accepted in their fields as means of conducting their inquiries. They must begin, at least, with folk psychological categories to describe what humans do when we introspect and metacognize (e.g., "we think about our thoughts"). And they only have recourse to reductive materialism or compatibilism as viable scientific accounts of consciousness. Consequently, they must also resort to folk psychological categories to explain nonhuman animal cognitive architecture.

As mentioned previously, we must start our exploration of consciousness somewhere and starting with Nagel's idea that consciousness is "how it feels to be in some state" seems as good a place as any (Nagel, 1974). There is something *it feels like* to be confused, hesitant, uncertain. However, this starting point has its shortcomings. To think about how it feels to be in some state requires that we represent our feelings to ourselves, something we are not particularly adept at doing. We are prone to confabulate, misunderstand ourselves, overestimate our virtues and underestimate our vices. We are especially good at "cooking the data" when looking in the mirror. We are subject to self-deception, taking ourselves either to be much better looking and more intelligent than we are, or much more pitiable and unlikeable than we are. We are rarely transparent to ourselves. Since our verbal self-reports must be taken with a grain of salt, any representation we make of our own consciousness should be cross-checked by the assessments of others. Only in this way will our representations of ourselves become more objective, less biased, and more truthful.

A word about what counts as higher and lower order thought. Morgan's canon cautions us not to attribute higher level psychological states to individuals if their behaviors can be explained in lower level terms. Several points about Morgan's canon merit attention. First, the canon assumes animals do have minds and undergo psychological processes. Second, it is anthropocentric. Morgan assumes that any explanation of animal behavior must begin with the "terms of the only mind of which we have first-hand knowledge," namely, our own (Morgan, 1894). Third, it does not simply call for the simplest explanation where 'simplest' refers to the entities (events, processes, properties, objects etc.) posited in reality, as with Occam's razor.[6] Fourth, it requires introspection on our own minds aimed at generating explanatory possibilities. Karin-D'Arcy (2005, p. 180) contends that "Morgan intended the canon to encourage comparative psychologists, through careful introspection, to attend to the levels of the functioning of human minds so that the activities of other species could be matched with the appropriate human functions, not those first intuited."

Morgan's canon tells us to introspect *critically*, identify the most general cognitive mechanisms necessary to explain the target behavior of an animal, and then apply the results in interpreting the observed animal behavior. While Carruthers (Carruthers, 2008) accepts Morgan's canon he does not believe that introspection is a reliable source of information about judgments and decisions or, more generally, most propositional attitudes (Carruthers, 2011). He accepts introspective knowledge of perceptual events, images, and emotions. So, if animal cognition is imagistic, as he thinks is a possibility (Carruthers, 2008, p. 64), then animals may have introspective knowledge of their states. Carruthers' theory leaves open the possibility that animals metacognize.

A further clarification. Since Morgan's canon begins with introspection, it might appear that Carruthers' explanatory schema is inconsistent with Morgan's canon. But this does not follow. Carruthers' explanatory schema does not rely on any controversial information obtained from introspection. He accepts the existence of folk psychological beliefs and desires and the ability, in humans

[6] Karin-D'Arcy (2005, p. 179) distinguishes Morgan's canon "from its intellectual predecessors," Occam's razor and Hamilton's law of parsimony, with which it is often erroneously conflated. The latter principles call for the simplest explanation in terms of entities posited in metaphysical investigations (Occam's razor) and in scientific-naturalistic investigations (Hamilton's law). Morgan's canon, at least, is a member of this family of concepts that includes Occam's razor and Hamilton's law.

and animals, for individuals to act in accord with their beliefs and desires. If we accept folk psychology, as I do, then we can accept introspection as a folk psychological jumping-off place.

But it is a dangerous jumping-off place. Our thoughts about our thoughts may be biased by overly generous assessments of the veracity of our introspective powers (Hurlburt, 2009; Hurlburt & Akhter, 2006; Winkielman & Schooler, 2012). If, as seems increasingly plausible, we confabulate narratives after the fact to explain almost all of our own actions, then we may not in fact be thinking about our own thoughts as often or as accurately as we think we are. If we have a non-metacognitive experience and, in retrospect, fool ourselves into believing that that experience was metacognitive, then the differences between our mental lives and the mental lives of animals may be more a matter of imagination than fact. For we may have far fewer metacognitive experiences than we suppose. And animals may have more metacognitive experiences than we assume.

How do we decide whether we are metacognizing when we think we are metacognizing? To answer this question we must proceed to the next step, designing a diagnostic tool capable of discriminating between genuine and illusory episodes of conscious metacognition in humans. Once that tool is in hand we can use it to decide whether any animals have conscious metacognition.

An Empirical Test for Conscious Metacognition in Humans

What empirical evidence would lead us to believe a human being's honest report that they were metacognizing at some point in time? Ideally, we would have cross-referenced multi-scale and multi-modal evidence of at least six kinds.

Step 1: Verbal Report

Verbal reports of the phenomenological presence of metacognition, however well-intentioned, cannot be trusted. However, subjects trained to screen off their biases and confabulations can learn to distinguish between events in which their beliefs are targeted at the world and events in which their beliefs are targeted at themselves. Given our propensity to tack *post-hoc* motives and intentions back onto our behaviors, we must approach self-reports of metacognition somewhat skeptically (Gazzaniga 2005; Wilson 2004). Given the fragmented and vacillating character of our conscious experiences, the experimental conditions necessary to produce reliable reports of metacognition require more elaboration than one might expect.

What would such conditions entail? We might take our clues from Russell Hurlburt. When he set out to discover what humans are thinking at random moments during the day, he found that he had to work with them for a period of weeks before they could accurately capture the contents of (their own!) consciousness. Over a period of years, Hurlburt devised the Descriptive Experience Sampling technique (DES). In DES, subjects are given a beeper that randomly emits a noise. When they hear it, subjects are to stop and write down notes about what was in their "pristine experience" at the moment of the sound (Hurlburt & Akhter, 2006). However, the procedure is not intuitive and subjects must be trained how to do it. Hurlburt educates them not to describe their experiences using phrases beginning "I was thinking about…" or "I was reading…" as these phrases typically point to the context or background of a subject's experience. Hurlburt wants to know the phenomena, the qualia, the directly observed objects of perception that are actually in "the footlights of consciousness" at the moment the beeper sounds. After training, subjects become skilled and begin to get it right, describing carefully what they experience.

Subjects should be carefully chosen for professional training in, and evidence of, metacognition. They might include, for example, psychologists and philosophers who do research on the subject, authors who have written critically received autobiographies, and meditation instructors who are finely attuned to their mental states. These experts would then be trained to engage in activities designed to elicit metacognition. They might be told first to record a specific memory, desire or aspiration. Then they might be instructed to record their thoughts about that memory or desire, explain how it arose, what level of

confidence they have in it, whether they think it is common or idiosyncratic, whether it was ephemeral or lasting.

Untrained reporters of their thoughts about their thoughts can mis-identify cognitive states as metacognitive because they do not accurately introspect themselves. They may, for example, draw on memories of their alleged metacognitive experience rather than the target experience itself (Hurlburt, 2009). This problem can be solved if something like Hurlburt's method is adopted. Unreliable reports in which the subject infers or presupposes that they were metacognizing can be screened off in favor of high-fidelity reports of actual metacognition.[7]

Screened self-reports from skilled metacognizers must also be time-stamped. That is, subjects must be assisted in reporting precisely when the target mental activity begins, when it ends, and what it contains. Metacognition is an expensive activity; the brain can support it for only so long. Any given subject, furthermore, is likely to move in and out of the desired activity, spending at times no longer than a few seconds in it and at others, perhaps in the case of a trained meditator proficient in Buddhist techniques, several hours. Subjects might be given stop watches and trained in using them to mark the beginning and ending of their meta-representing of themselves.

Step 2: Behavioral Profile

When, if ever, we have in hand trustworthy time-stamped self-reports of a subject's metacognizing, we may then obtain measurements of their physiological responses during those moments. At the whole-organism level, metacognition might be correlated, for example, with furrowed brows and lifted eyes, elevated heart rates, increased sweat production in the hands and cortisol levels in the blood. Here we would enlist medical assays to measure the responses of, for example, the cardiovascular, immune, and proinflammatory cytokine systems during the periods when subjects are genuinely thinking about their thoughts.

Step 3: Neural Correlates

At the next level we would establish fine-grained descriptions of what the brain is doing during metacognition. Suppose that Koch (2004) is correct that the neural correlates of consciousness (NCC) are "the *minimal set of neuronal events and mechanisms jointly sufficient for a specific conscious percept"* (p. 16). Suppose further that one is conscious if and only if the NCC undergo specific sequences of electrical activity, pulses, about a tenth of a volt in amplitude and 0.5–1.0 ms in duration (Koch, 2004). It is when this activity spikes, that is, reaches an action potential, that conscious awareness occurs. Finally, suppose that in the NCC are sub-networks of neural patterns that support awareness of one's mental states. There, we may hypothesize, lie the neural correlates of metacognition (NCM). For purposes of illustration, let us suppose that the NCM are stretched within and between areas of the precuneus and amygdala, the anterior and posterior cingulate cortex, the inferior parietal lobe, the ventromedial prefrontal cortex, and the hippocampus (Christoff, Gordon, Smallwood, Smith, & Schooler, 2009, Josipovic, 2014). If, to borrow a phrase from Searle (1994), cognition (whether "meta" or otherwise) must be tied to *inner, causally relevant* structures then, if Koch's hypothesis is correct, activation of the NCM would seem to be sufficient for metacognition.

Metacognition is by definition a higher-order process. Since higher-order states require the outputs of lower-order states, higher-order states are often temporally more extended and take more milliseconds to complete their operations. HOT states must fire later and longer than first-order states because they are composed of and supported by first-order states. If HOT theory is correct, episodes of

[7] Hurlburt (2009) writes that the results of his Descriptive Experimental Sampling psychology tests are consistent with Baars' global workspace theory and Carruthers' our-minds-are-opaque-to-us cognitive architecture.

metacognition last longer than first-order perceptual episodes. Since conscious perception can occur without activation of cerebral cortical areas (Merker, 2007), and since higher order thoughts, for all we know, require cortical involvement, then we can roughly characterize thoughts as either first- or second-order based on how long the thought persists in working memory. For a thought to be conscious it must, according to the theory, be globally broadcast, a process that takes no less than 0.4–0.5 s (Boly et al., 2013; Carruthers, 2011; Herzog et al., 2016). A neuro-anatomical test for metacognition, then, asks two questions. First, is a target mental process extinct before a third- to a half-second, a sign that it is probably not second-order? Second, if the target state lasts longer than 0.4–0.5 seconds, are there feedforward and feedback interactions between the cortex and subcortical networks during that time? Only in this case could the process be under the control of the agent. Only in this case could the state in question be a candidate for metacognition.

I should reiterate an important point, that we do not now know which parts of the brain are required for metacognition. Nor do we know how the feedforward and feedback loops behave as we think about our thoughts. Metacogition may well involve neural regions and cortical networks that we have not yet discovered. We need further research with humans, and the development of the technology necessary to support it, to inform our search for similarities and dissimilarities with animals.

Step 4: Cellular Processes

With reliable reports of metacognition correlated with whole-organism physiological responses and NCM in hand, we would then employ sub-organismic technologies to determine the cellular sponsors of metacognition.

All mammals have all of the brain structures named in the prior section on neural correlations. All mammals probably have, in addition, the capacity globally to broadcast the outputs of at least some submodules. However, while there are significant continuities between the neural networks of the brains of, say, monkeys and humans, there is not necessarily the same kind of continuity between the types of cells that make up the structures. This difference could make a difference to metacognition. I limit myself to one example.

Von Economo neurons (VENs), or spindle neurons, are large, uniquely shaped, and found in the frontoinsular cortex and anterior cingulate cortex (Allman et al., 2010). Whereas many neuronal cells have many dendrites, VENs have but one. For many years, scientists believed only humans had spindle cells, but we now know that all great apes have them (Nimchinsky, Vogt, Morrison, & Hof, 1995). So do many cetacean species and elephants (Marino et al., 2007). However, no evidence of VENs has been found in gibbons, monkeys, and prosimians (Nimchinsky et al., 1999). There may be other types of neural cells possessed by humans but not, say, by macaques. If so, and if unique kinds of cells are implicated in the NCM, or if the NCM have, say, higher concentrations of special cellular types than are found in the homologous brain areas of other species, then we might need to raise a flag about metacognition even if the prior four levels have been satisfied.

We need to know, then, about the composition and distribution of cell types, especially cell types that could be unique to humans, synaptic numbers and shapes, microcircuit connections between cells, and organization into layers in the cortex, in any species for which metacognition is proposed. Again, since we are uncertain about which kinds of cells are required for metacognition, or even if any particular kinds of cells are required, additional research is needed about this aspect of human metacognition in order properly to ground judgments about the presence or absence of nonhuman metacognition.

Step 5: Genetic Background

Just as there are significant continuities between the kinds of cells found in our brains and nonhuman primate brains, so are the majority of our genes similarly conserved—and famously conserved, as seen in the frequency with which we read that some variant of "humans share 98% of their genome with chimpanzees." For example, the so-called FoxP gene may be required for processes related to

metacognition. The FoxP gene is associated with learning and cognition in humans and is partly genetically responsible for a neural cluster involved in deliberation. It is found through much of the animal kingdom, including *Drosophila*, which show hesitation in some learning situations (DasGupta, Ferreira, & Miesenböck, 2014). That said, if some of the two percent of genes that we do not share with chimpanzees are involved in the NCM, then flags might again be raised.

We will want to know, then, about the composition and distribution of proteins, especially proteins that could be unique to humans, gene numbers and interactions, and up- and down-regulation of gene expressions in any species for which metacognition is proposed.

Step 6: Mathematical Information

Finally, if global workspace theory is compatible with an integrated information theory (IIT) of consciousness, then we should measure the amount and quality of information processed by the complex of elements constituting the NCMs. According to Tononi (2008), *integrated* information is:

> the amount of information generated by a complex of elements, above and beyond the information generated by its parts. Qualia space (Q) is a space where each axis represents a possible state of the complex, each point is a probability distribution of its states, and arrows between points represent the informational relationships among its elements generated by causal mechanisms (connections). Together, the set of informational relationships within a complex constitute a shape in Q that completely and univocally specifies a particular experience (p. 216).

IIT theory generates a mathematical measure of integrated information which Tononi calls measure "Φ." Φ is a quantified reduction of uncertainty (i.e., the information) that is generated when a system enters a particular state. Through causal interactions among the system parts above and beyond the information generated independently within the parts, information generates a range of conscious states, from weaker to stronger. If IIT is true, then metacognition requires multiple and complex informational relationships because stronger conscious states have higher numerical Φ values. Suppose that the scale of Φ runs from 0, a complete absence of conscious thought, to 1.0, the highest achievement of conscious thought. Suppose, further, that metacognition occurs around 0.9. Under these assumptions, the probability that an animal is metacognizing is higher if her Φ value reaches 0.8 than if her upper bound is 0.6. A computational test for metacognition would then ask two questions. What is the lower bound of Φ when any human metacognizes? And, does the Φ of any animal acting as if it is metacognizing rise to this level? If IIT is true and we eventually acquire an answer to the first question, then we would know which numerical values and which range of shapes to look for in the animal metacognition data (Tononi, 2008).

What empirical evidence would lead us to believe a human being's honest report that they were metacognizing at some point in time? If evidence from each of these six levels were found in a statistically significant number of human subjects and the data were cross-referenced at the appropriate scales and time periods, one would have a multi-modal scientific picture of metacognition.

Suppose our subject human cannot speak. A non-linguistic, aphasic, locked-in human would, by definition, be incapable of communicating their metacognitive acts. We would not be able to obtain self-reports from them, and so we would be stymied at step 5.1. Nonetheless, we would be on sound analogical grounds to infer metacognition in this subject were we to obtain positive evidence at the other six levels. In sum, the experimental method proposed here includes neuroanatomical, chemical, physical, and mathematical parameters of the metacognitive state. In the event that a subject lacks the linguistic capacity to tell us that they are metacognizing at a certain point in time, the method provides a way to help us determine whether they are.

If the global broadcast/HOT model is the right account of consciousness, and the tool just described can determine conscious metacognition in humans, the question of animal metacognition is two-fold. Does a given animal have a global broadcast/HOT system? And, if it does, does the system include representational contents that allow the animal to attend to its own cognitive states?

An Empirical Test for Conscious Metacognition in Animals

Suppose we have a profile in which honest time-stamped objective reports of metacognition in humans are correlated with the range of values at the physiological, behavioral, cellular, chemical, and mathematical levels named above. Suppose further that we have a nonhuman animal who is incapable of self-reporting metacognition and yet whose values in the other six steps fall within those of metacognizing humans. Since, by hypothesis the relevant activities and structures of NCM in humans have homologues in the animals in question, we will have a strong argument by analogy that the animals, too, are metacognizing. For example, spikes in an experimental monkey's Φ values during the time period of a trial would provide strong *prima facie* evidence that the monkey is metacognizing. The argument would go like this:

A. When engaged in metacognitive behavior, humans accurately self-report a state in step 1 with an empirical profile consisting of a suite of values determined experimentally by the method in steps 2 - 6.

B. Suppose that the state in question includes all of the following values: a blood cortisol level of between 24-28 mcg/dL; increased perspiration in the hands; activation of the precuneus, posterior cingulate cortex, inferior parietal lobe, medial prefrontal cortex, and hippocampus; heightened arousal of 100,000-120,000 von Economo neurons; EEG gamma frequency band activity of 38-44 Hz; and a Φ value between 9.0 and 1.0. Call any state satisfying all of these conditions ***M***.

C. Therefore, any animal with an empirical profile of ***M*** is metacognizing.

If the model proposed here is correct, having profile ***M*** suffices to qualify one as a metacognizer. The argument by analogy is straightforward. Assuming the present model, any animal who behaves, by our most sensitive assessments, *as if* metacognizing *and* who satisfies ***M*** *is*, on those grounds, metacognizing. And any animal who shows no behavioral indication of thinking about thinking and whose empirical profile contains values that do not satisfy ***M*** is, on those grounds, not metacognizing.

I have argued that inquiry into conscious metacognition must draw on many empirical disciplines--mathematics, genetics, cellular biology, neuroscience, and psychology—to construct a causal explanation of conscious metacognition in humans. This explanation, once achieved, can in turn provide a touchstone for claims about conscious metacognition in animals. As we do not yet have the required explanation of conscious human metacognition, however, we cannot fully justify claims that some animals engage in conscious metacognition. Should we eventually come into possession of the human explanation then we can conduct the required experiments with animals. And should those results eventually converge on a positive finding, we will at that point know that animals are conscious of what they know.

Acknowledgments

I am grateful to Bill Bauer, Mike Beran, and an anonymous reviewer for helpful suggestions, and to Bill for getting me started on the project.

References

Allman, J. M., Tetreault, N. A., Hakeem, A. Y., Manaye, K. F., Semendeferi, K., Erwin, J. M., … Hof, P. R. (2010). The von Economo neurons in frontoinsular and anterior cingulate cortex in great apes and humans. *Brain Structure and Function, 214,* 495–517. https://doi.org/10.1007/s00429-010-0254-0

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences, 6,* 47–52. https://doi.org/10.1016/S1364-6613(00)01819-2

Baars, B. J. (2003). Working memory requires conscious processes, not vice versa: A global workspace account. In N. Osaka (Ed.), *Neural basis of consciousness* (pp. 11–26). Philadelphia, PA: John Benjamins Pub Co.

Baars, B. J. (2005a). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience? *Progress in Brain Research, 150,* 45–53.

Baars, B. J. (2005b). Subjective experience is probably not limited to humans: The evidence from neurobiology and behavior. *Consciousness and Cognition, 14,* 7–21. https://doi.org/10.1016/j.concog.2004.11.002

Baars, B. J., & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences, 7,* 166–172. https://doi.org/10.1016/S1364-6613(03)00056-1

Basile, B. M., Schroeder, G. R., Brown, E. K., Templer, V. L., & Hampton, R. R. (2015). Evaluation of seven hypotheses for metamemory performance in rhesus monkeys. *Journal of Experimental Psychology: General, 144,* 85–102. https://doi.org/10.1037/xge0000031

Beran, M. J., Brandl, J., Perner, J., & Proust, J. (Eds.) (2012). *Foundations of metacognition*. Oxford, UK: Oxford University Press.

Beran, M. J., Smith, J. D., & Perdue, B. M. (2013). Language-trained chimpanzees (*Pan troglodytes*) name what they have seen but look first at what they have not seen. *Psychological Science, 24,* 660–666. https://doi.org/10.1177/0956797612458936

Beran, M. J., Smith, J. D., Redford, J. S., & Washburn, D. A. (2006). Rhesus macaques (*Macaca mulatta*) monitor uncertainty during numerosity judgments. *Journal of Experimental Psychology: Animal Behavior Processes, 32,* 111–119. https://doi.org/10.1037/0097-7403.32.2.111

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in Philosophy of Science, 9,* 261–325.

Boly, M., Seth, A. K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., … Tsuchiya, N. (2013). Consciousness in humans and non-human animals: Recent advances and future directions. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00625

Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. Cambridge, UK: Cambridge University Press.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language, 23,* 58–89. https://doi.org/10.1111/j.1468-0017.2007.00329.x

Carruthers, P. (2009a). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32,* 121–138. https://doi.org/10.1017/S0140525X09000545

Carruthers, P. (2009b). Mindreading underlies metacognition. *Behavioral and Brain Sciences, 32,* 164–182. https://doi.org/10.1017/S0140525X09000831

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford; New York: Oxford University Press, USA.

Carruthers, P. (2013). Animal minds are real, (distinctively) human minds are not. *American Philosophical Quarterly, 50,* 233–248.

Carruthers, P. (2014). Two concepts of metacognition. *Journal of Comparative Psychology, 128,* 138–139. https://doi.org/10.1037/a0033877

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106,* 8719–8724. https://doi.org/10.1073/pnas.0900234106

Church, B. A., Jackson, B. N., Beran, M. J., & Smith, J. D. (2019). Simultaneous versus prospective/retrospective uncertainty monitoring: The effect of response competition across cognitive levels. *Journal of Experimental Psychology: Animal Learning and Cognition,* 45, 311–321. https://doi.org/10.1037/xan0000207

Churchland, P. M., & Churchland, P. S. (1998). *On the contrary: Critical essays, 1987-1997*. Cambridge, MA: MIT Press.

Churchland, P. S. (1989). *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, MA: MIT Press.

Comstock, G., & Bauer, W. A. (2018). Getting it together: Psychological unity and deflationary accounts of animal metacognition. *Acta Analytica*, 33, 431-451. https://doi.org/10.1007/s12136-018-0340-0

Couchman, J. J., Coutinho, M. V. C., Beran, M. J., & Smith, J. D. (2009). Metacognition is prior. *Behavioral and Brain Sciences,* 32, 142–142. https://doi.org/10.1017/S0140525X09000594

Couchman, J. J., Coutinho, M. V. C., Beran, M. J., & Smith, J. D. (2010). Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition. *Journal of Comparative Psychology, 124,* 356–368. https://doi.org/10.1037/a0020129

DasGupta, S., Ferreira, C. H., & Miesenböck, G. (2014). FoxP influences the speed and accuracy of a perceptual decision in Drosophila. *Science, 344,* 901–904. https://doi.org/10.1126/science.1252114

DeGrazia, D. (2009). Self-awareness in animals. In R. W. Lurz (Ed.), *The philosophy of animal minds* (pp. 184–200). Cambridge, UK: Cambridge University Press.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience, 4,* 752–758. https://doi.org/10.1038/89551

Dennett, D. C. (1991). *Consciousness explained* (1st ed). Boston: Little, Brown and Co.

Dennett, D. C. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.

Dienes, Z., & Perner, J. (2002). The metacognitive implications of the implicit-explicit distinction. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 171–189).

Dretske, F. I. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.

Dutta, A., Shah, K., Silvanto, J., & Soto, D. (2014). Neural basis of non-conscious visual working memory. *NeuroImage, 91,* 336–343. http://dx.doi.org.prox.lib.ncsu.edu/10.1016/j.neuroimage.2014.01.016

Edelman, D. B., Baars, B. J., & Seth, A. K. (2005). Identifying hallmarks of consciousness in non-mammalian species. *Consciousness and Cognition, 14,* 169–187. https://doi.org/10.1016/j.concog.2004.09.001

Ferrigno, S., Kornell, N., & Cantlon, J. F. (2017). A metacognitive illusion in monkeys. *Proceedings of the Royal Society B: Biological Sciences, 284,* 20171541. https://doi.org/10.1098/rspb.2017.1541

Foote, A. L., & Crystal, J. D. (2007). Metacognition in the rat. *Current Biology, 17,* 551–555. https://doi.org/10.1016/j.cub.2007.01.061

Gallup, G. G. (1982). Self-awareness and the emergence of mind in primates. *American Journal of Primatology, 2,* 237–248. https://doi.org/10.1002/ajp.1350020302

Gazzaniga, M. S. (2005). *The ethical brain*. New York: Dana Press.

Gennaro, R. J. (2009). Animals, consciousness, and I-thoughts. In R. W. Lurz (Ed.), *The philosophy of animal minds* (pp. 184–200). Cambridge, UK: Cambridge University Press.

Goff, P., Seager, W., & Allen-Hermanson, S. (2017). Panpsychism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2017). Stanford, CA: Stanford University Metaphysics Research Lab.

Goldman, A. (1993). Consciousness, folk psychology, and cognitive science. *Consciousness and Cognition, 2,* 364–382. https://doi.org/10.1006/ccog.1993.1030

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative Cognition & Behavior Reviews, 4,* 17–28.

Hassin, R. R., Bargh, J. A., Engell, A. D., & McCulloch, K. C. (2009). Implicit working memory. *Consciousness and Cognition, 18,* 665–678. https://doi.org/10.1016/j.concog.2009.04.003

Herzog, M. H., Kammer, T., & Scharnowski, F. (2016). Time slices: What is the duration of a percept? *PLoS Biology, 14*. https://doi.org/10.1371/journal.pbio.1002433

Hurlburt, R. T. (2009). Unsymbolized thinking, sensory awareness, and mindreading. *Behavioral and Brain Sciences, 32,* 149–150. https://doi.org/10.1017/S0140525X09000673

Hurlburt, R. T., & Akhter, S. A. (2006). The descriptive experience sampling method. *Phenomenology and the Cognitive Sciences, 5,* 271–301. https://doi.org/10.1007/s11097-006-9024-0

Josipovic, Z. (2014). Neural correlates of nondual awareness in meditation. *Annals of the New York Academy of Sciences, 1307,* 9–18. https://doi.org/10.1111/nyas.12261

Jozefowiez, J., Staddon, J. E. R., & Cerutti, D. T. (2009). Metacognition in animals: How do we know that they know? *Comparative Cognition & Behavior Reviews, 4,* 29-39. https://doi.org/10.3819/ccbr.2009.40003

Karin-D'Arcy, M. R. (2005). The modern role of Morgan's canon in comparative psychology. *International Journal of Comparative Psychology, 18.*

Koch, C. (2004). *The quest for consciousness: A neurobiological approach* (1st ed.). Denver, CO: Roberts & Company Publishers.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9,* 149–171. https://doi.org/10.1006/ccog.2000.0433

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289–326). Cambridge, UK: Cambridge University Press.

Kornell, N. (2009). Metacognition in humans and animals. *Current Directions in Psychological Science, 18,* 11–15. https://doi.org/10.1111/j.1467-8721.2009.01597.x

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 686–708. https://doi.org/10.1037/a0026478

Lewis, D. (1966). An argument for the identity theory. *The Journal of Philosophy, 63,* 17–25. https://doi.org/10.2307/2024524

Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy, 67,* 427–446. https://doi.org/10.2307/2023861

Lycan, W. G. (1987). *Consciousness*. Cambridge, MA: MIT Press.

Lycan, W. G. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21,* 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Marino, L., Connor, R. C., Fordyce, R. E., Herman, L. M., Hof, P. R., Lefebvre, L., … Whitehead, H. (2007). Cetaceans have complex brains for complex cognition. *PLoS Biology, 5*. https://doi.org/10.1371/journal.pbio.0050139

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30,* 63–81. https://doi.org/10.1017/S0140525X07000891

Morgan, C. L. (1894). An introduction to comparative psychology. London: Walter Scott.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83,* 435–450.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51,* 102–116. https://doi.org/10.1037/0003-066X.51.2.102

Nimchinsky, E. A., Gilissen, E., Allman, J. M., Perl, D. P., Erwin, J. M., & Hof, P. R. (1999). A neuronal morphologic type unique to humans and great apes. *Proceedings of the National Academy of Sciences of the United States of America, 96,* 5268–5273.

Nimchinsky, E. A., Vogt, B. A., Morrison, J. H., & Hof, P. R. (1995). Spindle neurons of the human anterior cingulate cortex. *The Journal of Comparative Neurology, 355,* 27–37. https://doi.org/10.1002/cne.903550106

Oakley, D. A., & Halligan, P. W. (2017). Chasing the rainbow: The non-conscious nature of being. *Frontiers in Psychology, 8*, 1-16. https://doi.org/10.3389/fpsyg.2017.01924

Penrose, R. (1994). *Shadows of the mind: A search for the missing science of consciousness*. Oxford: Oxford University Press.

Penrose, R. (2016). *The emperor's new mind: Concerning computers, minds and the laws of physics* (Revised impression as Oxford landmark science). Oxford: Oxford University Press.

Proust, J. (2009). Overlooking metacognitive experience. *Behavioral and Brain Sciences, 32,* 158–159.

Proust, J. (2010). Metacognition. *Philosophy Compass, 5,* 989–998. https://doi.org/10.1111/j.1747-9991.2010.00340.x

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. Oxford: Oxford University Press.

Ravenscroft, I. (2019). Folk psychology as a theory. In E. N. Zalta (Series Ed.), *The Stanford encyclopedia of philosophy*. Stanford, CA: Stanford University Metaphysics Research Lab.

Rosati, A. G., & Santos, L. R. (2016). Spontaneous metacognition in rhesus monkeys. *Psychological Science, 27,* 1181–1191. https://doi.org/10.1177/0956797616653737

Rosenberg, A. (2011). *The atheist's guide to reality: Enjoying life without illusions* (1st ed.). New York: W.W. Norton.

Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford: Oxford University Press.

Rowlands, M. (2001). Consciousness and higher-order thoughts. *Mind & Language, 16,* 290–310. https://doi.org/10.1111/1468-0017.00171

Searle, J. R. (1994). Animal minds. Midwest *Studies in Philosophy, 19,* 206–219.

Shea, N. (in press). Concept-metacognition. *Mind & Language.*

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences, 18,* 186–193. https://doi.org/10.1016/j.tics.2014.01.006

Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review, 68,* 141–156. https://doi.org/10.2307/2182164

Smith, J. D. (2005). Studies of uncertainty monitoring and meta-cognition in animals and humans. In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition: Origins of self-reflective consciousness* (pp. 242–271). Oxford: Oxford University Press.

Smith, J. D., Beran, M. J., Couchman, J. J., & Coutinho, M. V. C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review, 15,* 679–691. https://doi.org/10.3758/PBR.15.4.679

Smith, J. D., Beran, M. J., Redford, J. S., & Washburn, D. A. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General, 135,* 282–297. https://doi.org/10.1037/0096-3445.135.2.282

Smith, J. D., Church, B. A., Beran, M. J., & Washburn, D. A. (2018). Meta-cognition. In J. Vonk & T. Shackelford (Eds.), *Encyclopedia of animal cognition and behavior*. Basel, Switzerland: Springer.

Smith, J. D., Couchman, J. J., & Beran, M. J. (2014). Animal metacognition: A tale of two comparative psychologies. *Journal of Comparative Psychology, 128,* 115–131. https://doi.org/10.1037/a0033105

Smith, J. D., Coutinho, M. V. C., Church, B. A., & Beran, M. J. (2013). Executive-attentional uncertainty responses by rhesus macaques (*Macaca mulatta*). *Journal of Experimental Psychology: General, 142,* 458–475. https://doi.org/10.1037/a0029601

Smith, J. D., Redford, J. S., Beran, M. J., & Washburn, D. A. (2010). Rhesus monkeys (*Macaca mulatta*) adaptively monitor uncertainty while multi-tasking. *Animal Cognition, 13,* 93–101. https://doi.org/10.1007/s10071-009-0249-5

Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences, 26,* 317–373.

Smith, J. D., & Washburn, D. A. (2005). Uncertainty monitoring and metacognition by animals. Current *Directions in Psychological Science, 14,* 19–24. https://doi.org/10.1111/j.0963-7214.2005.00327.x

Stoerig, P., & Cowey, A. (1995). Visual perception and phenomenal consciousness. *Behavioural Brain Research, 71,* 147–156.

Strawson, G. (2006). Realistic materialism: Why physicalism entails panpsychism. *Journal of Consciousness Studies, 13,* 3–31. https://doi.org/10.1093/acprof:oso/9780199267422.003.0003

Tegmark, M. (2000). Importance of quantum decoherence in brain processes. *Physical Review E, 61,* 4194–4206. https://doi.org/10.1103/PhysRevE.61.4194

Tolman, E. C. (1938). The determiners of behavior at a choice point. *Psychological Review, 45,* 1–41. https://doi.org/10.1037/h0062733

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5,* 42. https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin, 215,* 216–242. https://doi.org/10.2307/25470707

Tye, M. (2003). A theory of phenomenal concepts. *Royal Institute of Philosophy Supplements, 53,* 91–105. https://doi.org/10.1017/S1358246100008286

Van Gulick, R. (2018). Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018). Stanford, CA: Stanford University Metaphysics Research Lab.

Vandekerckhove, M., & Panksepp, J. (2009). The flow of anoetic to noetic and autonoetic consciousness: A vision of unknowing (anoetic) and knowing (noetic) consciousness in the remembrance of things past and imagined futures. *Consciousness and Cognition, 18,* 1018–1028. https://doi.org/10.1016/j.concog.2009.08.002

Wilson, T. D. (2004). *Strangers to ourselves: Discovering the adaptive unconscious* (New ed.). Cambridge, MA: Belknap Press.

Winkielman, P., & Schooler, J. W. (2012). Consciousness, metacognition, and the unconscious. In S. T. Fiske & C. N. Macrae (Eds.), *The SAGE Handbook of Social Cognition* (pp. 54–74). Los Angeles, CA: SAGE Publishers.