Evolution of Metacognition

Janet Metcalfe

Columbia University

Please send correspondence to:

Janet Metcalfe

Department of Psychology

Columbia University

NY NY 10027

212-854-7971

jm348@columbia.edu

Abstract

The importance of metacognition, in the evolution of human consciousness, has been emphasized by thinkers going back hundreds of years. While it is clear that people have metacognition, even when it is strictly defined as it is here, whether any other animals share this capability is the topic of this chapter. The empirical data on non-human metacognition are reviewed. It is concluded that three monkeys have now shown evidence of metacognition. Even in these primates, however, the capabilities are limited. Despite claims that rats have metacognition, the data can be explained in terms of mere conditioning contingencies. No other species have been shown to have metacognition. Thus, metacognition appears to be a very recently evolved capability. It is one that may confer on humans an ability to escape from being stimulus bound, and allow self control of their learning, and actions.

Even before psychology was recognized as a separate discipline, scholars were fascinated by what we now call metacognition, because self-reflective knowledge (i.e., metacognition) was thought to embody a particular kind of consciousness unique to human beings. According to a number of thinkers, this kind of consciousness bears a special connection to our 'self' or our knowledge of ourselves, as in the maxim, 'know thyself.' The notion that there is a looker, embedded within our cognitive fabric, that is somehow able to look at our other cognitive processes, has such compelling force as being a special entity to have provoked early philosophers from St. Augustine (see, Harrison, 2006) to Descartes (1637/1999) to suppose that there is a disembodied soul. The modern analogue, while disavowing a non-physical soul, is to claim that this self-reflective capability is, nevertheless, a special mental capability and a phenomenological experience which is specific to humans. This view has been articulately espoused by moderns from Armstrong (l968) to Rosenthal (2002), and holds considerable appeal. The idea is that whereas other species may have evolved adaptive characteristics such as the ability to fly, or, like the raptors, to see tiny movements many miles away, or, like the monarch butterfly, to eat foods that are poisonous to other animals, the human species has evolved--as its unique adaptive strength-- a particular form of consciousness. The most elementary component of this form of consciousness is metacognition.

Is metacognition a special kind of consciousness ?

Descartes, in what we now consider to be elaborate metacognitive musings, reached the conclusion that the fact of these musings--that he was able to think about his thinking-- gave indisputable proof of his own existence. What Descartes was doing,

when he was isolated in his poele (a small cabin with a woodstove) thinking about the basis of all knowledge, was deeply metacognitive. He was considering whether his physical body might be different, and he acknowledged that it might. He was thinking about whether his perceptions might be faulty—which all modern psychologists and an entire tradition focused on illusions and distortions and biases of perception –see, e.g., Hochberg, 2003--resonate to. He was deliberating over whether his memories of his own personal experience might be wrong. The vulnerability of memory is, of course, now well established, Loftus, 2004. Despite all these possibilities of cognitive and perceptual distortions—which we now know extend even to the metacognitions themselves (see Bjork, 1994; Jacoby, Bjork, & Kelley, l994; Metcalfe, 1998), what Descartes was unable to deny (c.f., Russell, l945/l972) was that there was somebody doing all of this reflection—him. This observation, that such metacognitive musings implicated a self who is the muser, had deep significance for Descartes, and for subsequent thinkers.

Descartes reached a conclusion that most modern neuroscientists (e.g., Damasio, 1994), even those who ascribe to the importance of metacognition as entailing a special state of consciousness, might shy away from, namely, that the existence of such self reflection implies that there must be a non-physical soul. Descartes, of course, was a dualist, and used his meditations to that end. However, one need not take a dualist stance to acknowledge the special status of metacognition in determining a particular kind of consciousness that may be available to humans and perhaps to other animals. The possible extension of this kind of consciousness to nonhumans was explicitly denied by Descartes, who believed that it, and hence the possibility of a soul, existed only in

humans. The primary evidence weighing in on Descartes' conclusion was that animals did not have language. And, to this day, though there have been many studies attempting to demonstrate that at least some non-human primates have language, none have done so definitively (Terrace, 2005; Terrace & Metcalfe, 2005).

To the non-dualist, who might nevertheless acknowledge self-reflective consciousness as a unique cognitive capability, it seems plausible that this special kind of consciousness may have arisen during the course of evolution, and it may have had as particular adaptive value for the animals who have it, namely us. It may allow them to do things (for example, to reflect on their actions and their outcomes, and change those actions as indicated by the reflection to obtain better results) that other animals cannot do. This ability to gain reflective control over their own behaviors may well have allowed our ancestors to survive under circumstances fatal to other animals. The advantages of being able to foresee and evaluate events in one's mind's eye beforehand rather than having one's actions driven solely by the afferent stimuli seems self-evident. Being able to reflect on past occurrences also has its own adaptive value, freeing such an animal from the constraints of the stimulus, and allowing more rational, adaptive future responding. Such consciousness may also have a benefit, to those who had it, in terms of sexual selection—its presence being particularly attractive to potential mates. Being able to take another's point of view--a sophisticated kind of metacognition known as theory of mind (Frith & Happe, l999; Heyes, l998; Leslie, l987; Perner, l991; Povinelli, 2000)-- is indisputably appealing. People like feeling understood. It could also allow the person who has this ability to deceive more effectively--a trait that while despicable might provide certain evolutionary advantages for the person who has it (see, Byrne & Whitten,

l992; deWaal, 1992;Whitten & Byrne, l988, for anecdotes about the deceptive behavior of non-human primates, and the consequences for mating success). One can entertain the idea that such a special kind of consciousness could evolve without necessarily accepting the postulate of Descartes that its existence is proof positive against materialism.

Comte's paradox

The introspection that there is inside of us some special-status looker who can observe its own internal cognitions resurfaced, in the last century, as Comte's paradox. A paradox is defined as an apparently true statement that leads to a contradiction or to a situation that defies intuition. For Comte, how the mind or consciousness could both function and observe itself function seemed paradoxical. The fact that metacognition was, until very recently, perceived as a paradox is based on the deeply felt idea that consciousness is unitary and indivisible, rather than piecemeal and fragmentary. The paradox depends on the statement being truly self-referential, in the strictest sense. But, as many perceptual psychologists have demonstrated, (see, Hochberg, 2003) perception is, itself, piecemeal and fragmentary, even though there is an illusion of a continuous whole. Perhaps the most dramatic example of this comes from recent change blindness (Simons & Chabris, 1999) studies, in which a person can be, for example, watching a videotape of a game of catch among several players, and appear to have a whole and continuous perception of the entire field, with all of the players in this field. But this apparent wholeness and continuity is belied by the fact that a full sized person in a gorilla costume walks through the scene, stopping to beat his chest in the middle of the screen, and people, watching the ball throwing, do not see it. When told about the gorilla, and shown the video again, they see it clearly, of course. Despite this gross omission--an

enormous blind spot-- they had no notion that there were any holes in their consciousness. It is simply that the notion of the unity of consciousness, and its apparent wholeness, is illusory. Our illusion of perceptual continuity (see Hochberg, 2003) is constructed from what we see and hear, from what we expect, and, in a fragmentary way, from what we see, with all of these components and a number of different modalities contributing in parallel.

Across modalities, it is straightforward to follow more than one line of consciousness, of course (so, cross modal monitoring would not be paradoxical). One can drive and listen to the radio at the same time, being aware of both. But even within a single modality, it has now been shown that the 'spotlight of attention' (Triesman, 1985), which was originally thought to be a single indivisible spotlight (as would be consistent with the idea that Comte's paradox might really be paradoxical) can be divided into two different, and spatially discontinuous locations (Hillyard, 2002), at the same time. Thus, as many elegant experimental studies of perception have shown, the assumption of a unitary consciousness does not hold up.

Furthermore, even if consciousness were unitary in each moment of psychological time, the possibility remains that 'function' and the reflection do not, in fact, co-occur in the same psychological moment. We might be able to observe our own mental function by taking a snapshot of it in one moment, and looking at that snapshot (or its ghost in working memory) in the next--alternating back and forth. Many studies of working memory illustrate this capability.

Finally, there is no contradiction of logic that people might be conscious of more than one thing at a time, simultaneously entertaining the cognition or memory and one's

assessment of it, in parallel. For Comte's paradox to be a paradox, and self-referential, the object reflected and the reflector must really be one and the same entity. From a neuroscience perspective, though, the brain is constantly monitoring and feeding back information at all levels. For example, Oschner and Gross (2006) have elaborated how the prefrontal cortex and the cingulate control system work in concert with subcortical (especially amygdala) emotional-generative systems, to allow the modulation of emotional responses. Attentional regulation directs and controls other cognitive processes, and different aspects interact in a complex manner, as has been illustrated by a meta-analysis conducted by Wager and Smith (2003). To suppose that this could not be so--that doing and monitoring, or functioning and observing the functioning, could not co-occur--might well be considered quaint by modern neuroscience criteria. Thus, for Comte's paradox to be a puzzle one must affirm as unassailable certain assumptions about consciousness and about brain function--assumptions that modern research refutes.

Even so, the postulation of a 'paradox' was taken seriously enough by early experimental researchers in metacognition to provoke an explicit theoretical solution. Nelson and Narens, l990, in response to this supposed conundrum, proposed that in order to allow that the mind could both function cognitively, and observe its own cognitive functions, that there must exist <u>two</u> levels (of consciousness), a base, or object level and a metalevel. This solution, of course, says that consciousness is not unitary--just as much modern neuroscience would affirm. This framework has been widely accepted.

Does metacognition imply an infinite regress?

The idea that there is a looker of sorts, functioning at the metalevel in Nelson and Narens' framework, also withstands the 'turtles all the way down', or infinite regress,

criticism. The criticism is based on the idea that if one has to have observation of cognition, then there must be a conscious observer inside the person's head. That observer needs to be able to see what is going on at the basic cognitive level, and so it needs to be a full blown internal person, or homunculus, complete with a fully elaborated perceptual-cognitive apparatus. But then one needs to propose that there is an homunculus inside the head of the homunculus, to be conscious of what it is seeing, and so on ad infinitum. This dissolves into absurdity. The 'turtles' criticism depends on the postulate that observation, or monitoring, entails an elaborate observer--essentially a full blown person. But monitoring, computationally, at least, can be extremely simple. A simple thermostat monitors the room temperature, and can trigger an action (turn off the heat) without anything like a full blown cognitive-perceptual apparatus. A model of metacognitive monitoring, that is sufficient to produce the kind of metacognitive data people give in feeling of knowing experiments, may involve only simple computation -- see Metcalfe, l993, who, within the CHARM framework, was able to model nearly all of the known data on the feeling of knowing phenomenon by postulating only a simple computation of a correlation between an input vector and a trace vector. This entails only one computation, and it is one that is well documented as existing in the nervous system . Certainly, then, the possibility of metacognition--if it entails only such straightforward computations-- is not threatened by the 'turtles all the way down' criticism.

It is interesting to note that it was not until our modern familiarity with ideas like semi-modular brain function, and parallel distributed cognitive processing capabilities, and a systems approach to the mind-brain, that researchers were able to free themselves of the idea that a self-reflective capability was a deeply perplexing paradox. We now

find the puzzlement puzzling, and agree with Humphrey (l987) in saying: " The problem of self-observation producing an infinite regress is, I think, phony. No one would say that a person cannot use his own eyes to observe his own feet. No one would say, moreover, that he cannot use his own eyes, with the aid of a mirror, to observe his own eyes. Then why should anyone say a person cannot, at least in principle, use his own brain to observe his own brain? "(p. 11).

Although we no longer view humans' metacognitive capability as either a paradox, or as bearing some kind of mystical meaning, we do not rule out the possibility that this particular capability may be unique to humans, or that it bestows on them some cognitive, and adaptive, capabilities that may be missing in other creatures. Despite being demystified, it may still be special. But to determine whether it is indeed specific to humans, and to investigate empirically this question, we need, first, to define what is meant by metacognitive monitoring and control.

Definition of Metacognition

There is monitoring and control at all levels of the human and the animal mind/brain system. Indeed, the entire brain can be thought of as a giant feedback system, with virtually every pathway having both feedforward and feedback connections, and multiple connections among different brain regions serving to allow the outcomes of one kind of processing to modulate other processes. So, if monitoring and feedback were all that was meant by metacognition it would be pervasive and there would be no question at all that most other animals also use such feedback. But it is not simple feedback from one level interacting with processing at another that, alone, characterizes metacognition.

Furthermore, it is not simply being able to make a discrimination or a judgment. Even very simple animals are able to make discrimination judgments about events in the world. Indeed, even non-animals can make some of these. A plant apparently 'judges' the lightness in its environment, and moves, very slowly, towards the light. Among animals, judgments about things in the world can be much more complex. A pigeon can make line length discriminations. A rat can make at least 8-alternative discriminations, and reliably take the correct arm of a radial maze. Many animals can make duration discriminations. And animals can show differential responses, including severe anxiety, when discriminations become very difficult. Pavlov (1927) made a circle a conditioned response to feeding, and an oval was made a food-negative response. Whenever a circle appeared the dog would get food. When an oval appeared it would not be fed. The poor dogs who, after this training, were exposed to stimuli half way between the ovals and the circles showed symptoms of severe anxiety. Tolman (1932), too, showed that animals given choices of stimuli between two discriminable categories, can be 'caught at the choice point,' being tugged simultaneously in two directions. The anxiety of Pavlov's dogs suggests that such conflict may well have visceral (and noticeable) consequences. But even such dramatic responding to very difficult discriminations do not qualify as being metacognition, since they are merely responses to the afferent stimuli, and do not concern judgments about internal representations.

Furthermore, the responses animals make can be quite complex, without making them qualify as metacognition. Circus trainers are able to get animals--through well understood conditioning techniques--to exhibit behaviors that are both complex, that are not seen in the animals in their normal untrained repertoire, and that may involve

multiple steps. This training, typically, starts with a simple response (perhaps as insignificant as getting the animal to turn in a certain direction, or move a certain way) and through many trials, builds on those initial small responses until an elaborate sequence of moves--like getting an elephant to stand on one foot on a bucket--can be produced. Thus, through this kind of shaping, animals can be trained to make fine-grained non-binary discriminations about what they see and hear in the world, and they can perform multiple step and complex responses. None of this requires metacognition.

Metacognition, then, is not merely a judgment among options, however refined, and regardless of the number of discriminanda. It is not merely the production of a complex multi-step response, to get a reward. And it is not the combination of a multi-step response to a difficult discriminative judgment. Instead, it is a very special kind of judgment or commentary that involves a level of processing that we, here, call representational, or cognitive (and that Nelson & Narens, l990, 1994, called the object level) and a higher level monitoring that we call metacognitive. A simple case of a cognition or a representation is a word, or a symbol. A word is not the object in the world itself, but rather it refers to the object, and is about the object. A memory is also a representation. It is not present in the world, but rather it is internal. If a memory is represented internally, and a person makes a judgment about that memory, then that judgment is a metacognitive judgment. Note, however, that judgments in some recognition tasks, in which the probes are given in the testing environment, do not qualify as being metacognitive, since the person can make the judgment based on the probe that is present in the afferent environment, and not the memory to which the probe refers. The probe, being present in the stimuli environment, is not properly considered to be a mental

representation, even if its ongoing processing has been influenced by something that happened in the past. (Note that this critique applies to virtually all implicit memory tasks. They are not metacognitive, by the present criterion.) If a person just makes a judgment about something that they see or hear, or even about their own current fluency of processing, it is not metacognitive, since it is not a judgment about a mental representation. Metacognition must be a judgment about an internal representation. Metacognition differs from mere judgment insofar as it is not stimulus bound, or directly related to something in the animal's afferent environment. Rather, it is about a mental representation. While denying metacognition, so defined, is supernatural, we might still maintain that it could be a truly extraordinary capability, and explore its implications and evolution.

Usually metacognition requires language (as Descartes intuited). The individual is asked whether or not they will know the answer to a question. To be unequivocal that the cognition being queried is representational, a question can be posed about something that is not present in the immediate environment, like a memory. The participant then gives a rating on some scale about the answer, or about whether they will be able to retrieve the answer later, for example. The question and the answer to the question are indisputably mental representations, or concepts at a cognitive level, so the rating is true metacognition. Although language is typically used in these assessments, if a researcher were clever enough to be able to administer metacognitive tests that were about non-verbal internal representations using responses such as betting, rather than, say, verbally-based rating scales, then it should be possible to determine whether animals have metacognition. And, indeed, there have been several recent attempts to do just that.

Do other primates have metacognition?

The attempt to determine whether any non-humans have metacognition is important for a number of reasons, not the least of which is the question of whether we can use an animal model to gain understanding of human thought. While nobody would dispute that animal models of human responding hold huge promise in some domains, such as pain, fear, and stress reactions, there may be distinct limits. If no animals other than humans have metacognition, then certain states of consciousness simply cannot be studied with any subject other than a human one. But perhaps animals have metacognition.

Call and Carpenter (2001) were among the first researchers to systematically attempt to investigate whether any non-humans have metacognition. They asked whether there was any evidence that great apes knew what they themselves knew. The paradigm that they used was clever. They showed chimps or orangutans a choice food morsel being hidden in one of two tubes. The apes reached immediately into the appropriate tube for the food. But then the researcher placed a barrier between his hand, hiding the food in one of two tubes, and the line of sight of the ape. The apes, in this condition, did not know where the food was hidden. The question they asked was: do the apes seek information when they know they do not know where the food is hidden? If they seek information, by looking into the tubes, before reaching, Call argued that this gives evidence that they know that they do not know, and that knowing that one does or does not know is metacognition. The looking behavior of the great apes was much greater in the situation in which the hiding was hidden than when it was exposed. Young children

of 2 years of age performed in much the same way as did the apes. But dogs, in contrast, did not seek information first (see Call, in Terrace & Metcalfe, 2005).

Is this metacognition? The basic tenant in this research is that information seeking indicates metacognition. This is an interesting perspective on the question, but one that deserves intensive scrutiny. Does moving one's eyes before reaching for an apple imply that one is using metacognition? If one found, for example, that squirrels or chipmunks or birds looked around --scanning the skies with their eyes or listening carefully with their ears for predators-- before venturing out on an open field, would one thereby grant them metacognition? If an animal were running on a rough pathway, or swinging through the jungle through the trees, would looking first before stepping or leaping, to see whether there was a hole at the next step, or whether the branch was thick or thin that they were going to grasp onto, be an indication of metacognition? Probably not.

Other researchers have investigated the possibility of metacognition in animals other than humans as well. Smith, Shields, and Washburn (2004) reviewed a series of experiments, mostly from their own labs, investigating the possibility of metacognition with apes, monkeys and dolphins. They likened metacognition to uncertainty judgments, or, for those not willing to say that non-humans are really 'judging,' to indications of uncertainty. So, if the animal gave evidence that it was not sure of the answer or of the course of action to follow, then this was taken by Smith and colleagues to be evidence for metacognition. It is interesting that Smith appears to have picked up on a different aspect of Descartes' thinking--the ability to doubt-- rather than the more standard self-reflective component.

Smith and colleagues conducted many classification tasks with animals, in which they were trained to make one response when to a particular category and a different response to a second category, on the same dimension. Then they would expose the animal to a situation in which the two categories blended smoothly into one another. An example would be a dot density discrimination task, where the animals were trained to make a response A to dense displays and a response B to less dense displays. They were then given displays of intermediate density. They allowed the animals to give an escape response to get some reward reliably, and found that in these intermediate or what they called 'don't know' situations, the animal would often choose to hit the escape button. These 'uncertainty' responses held along a number of dimensions, such a loudness, length of sound, pitch discrimination, or density, etc. They also held for a number of species--apes, monkeys and dolphins.

Furthermore, Smith, Shields, and Washburn (2005) have shown that the uncertainty functions in these animals have much the same form as did analogous functions when humans were the participants. Undoubtedly, humans and non human animals respond in a similar way on these materials. The question remains, though, as to whether these results indicate metacognition, either in the nonhumans or in the humans?

On several grounds, I suggest that the answer is 'no.' First, it is not obvious that the escape button really does mean to the animal that he does not know (even if it does have that meaning to the human). Maybe it just means that there is a third category--intermediate length lines or moderate density--for which it can get the best possible rewards by hitting the button that the experimenter thinks is the escape or uncertainty button. But to the animal this button is just a third category label. There is no question

that even animals less intelligent than dolphins can make at least eight item discriminations, witness the eight-arm radial maze used universally in studies with rats. So showing that a non-human animal can make a three-part, rather than just a binary discrimination is not evidence for metacognition.

Second, the stimuli about which the animals are responding are present in the environment that the animal can see, hear, smell or touch, when they start to make their responses, in these studies. They are not memories. Thus, even if the responses they are making are judgments (but see above), because they are not about internal representations, they are not metacognitive judgments. The elementary qualification that metacognition be a judgment about a representation is not met.

It is interesting that Smith et al. (2004) note in their review paper that it had been recommended by early researchers that the judgments animals make be done retrospectively--allowing them to give the primary response *then* make their confidence judgment, as is usually done with humans. This procedure would increase the chance that the judgment was about a representation rather than about the stimulus itself. But they note that "The catch is that animals have so far not been able to report their confidence this way" (p. 8). Because these studies do not meet this fundamental criterion of being about a representation, it seems prudent to be skeptical about whether any of these studies indicate metacognition in non-humans.

Hampton (2001, 2004), however, devised a task that, while not involving long-term memory, did involve an elementary form of memory. In an experiment with two rhesus monkeys, Hampton (2001) used a task called a delayed-match-to-sample task, in which the stimulus was no longer present in the environment when the monkeys had to

make a decision about whether or not to take a test. Thus, Hampton's paradigm goes a long way towards countering criticisms of Smith's procedures. The stimulus being judged was not present, so there was at least the possibility that the judgment was about an internal representation rather than about a stimulus that was present at the time of judgment. Furthermore, Hampton rotated through 4 stimuli each day, randomly choosing one of the four as the target on each trial. The monkey had to remember which stimulus was correct on each trial, and all four of the alternatives had been equally reinforced in this role. Thus, it was not merely a discrimination conditioning task (as could have been the case in the studies Smith reviewed). Thus, Hampton's task was a difficult memory task, not merely a conditioned discrimination in which the stimulus mappings were always the same.

On each session, Hampton presented the monkeys one of four images, that it had to touch on the computer touch-screen three times. This multiple touching was designed to improve the chances that the monkey saw the to-be-remembered item. Then, a delay was intervened, during which, on two thirds of the trials, the animal was given a choice of whether it wanted to take the test or decline to take the test. If it wanted to take the test, it touched one icon; to decline it touched another icon. If the monkey chose to take the test, it was given a 4-alternative forced choice test, with all four of the stimuli that were being used on that session as the alternatives, a few moments later. If it touched the item that it had seen on the present trial, it got a peanut. If it touched one of the three incorrect items, it got nothing. If the monkey declined to take the test, it got a primate pellet (which it liked more than nothing, but not as much as peanuts). On the remaining

one third of the trials, the monkey was forced to take the test, without an intervening choice.

The data on the first experiment showed that accuracy was better, for both monkeys, when they had chosen to take the test than when they had been forced to take the test. In an additional experiment, a time delay was manipulated. Though both monkeys chose to take the test more often at short intervals, and both monkeys numerically showed better performance when they chose, at all time intervals, the data for only one monkey showed this difference in performance to be significant.

Does this study show that monkeys have metacognition? First, since only one of the two monkeys showed a significant effect on all criteria, we might, at best, have evidence that one monkey has shown metacognition. Experimental psychologists testing humans, though, prefer larger sample sizes and more consistency, before reaching important conclusions, and would prefer a criterion of something like 1/20 that their results are not just an accident. Still, the second monkey did show effects in the right direction. Second, the delays in the match-to-sample task were rather short (at the longest only 240 seconds) relative to those used in some metacognitive studies with humans. Thus, it may be controversial that these working memory representations should really be considered memories, rather than something more akin to afterimages. Still, the stimulus itself was <u>not</u> present at the time the judgment was made, and this is a great improvement in methodology. Third, the task was not a simple discrimination learning task, but involved an ongoing and changing memory (albeit with a brief delay) and so the experiment avoids this criticism. And finally, the alternatives were not present when the judgment was made, so the judgment could not be made by simply assessing the fluency

of each alternative. (When the test questions are present, even pigeons can do such tasks.) The fact that the alternatives were not present when the judgment was made allows this experiment to avoid another criticism. These data, then, suggest--though perhaps not as strongly as one would like-- that monkeys may have some metacognitive capabilities. It was the first to do so.

Son and Kornell (2005) have also provided some data indicating that rhesus monkeys have at least a glimmering of metacognition. They trained two monkeys--Lashley and Ebbinghaus--to do a line length discrimination task. After the monkeys had seen the lines and made their choice of which was the longer (or shorter) line, they were then trained to select, on a touch screen, whether they wanted 'to bet' on their answer or not. Note that neither the stimulus nor their choice on the test was present on the screen (though there was no extended time interval between the response and the judgments. Note that this paradigm fits what early researchers had suggested and Smith et al., 2005, had thought could not be done). If the monkeys chose the 'high risk' icon on the touch screen, and their response had been correct, they received several token rewards which -- when enough tokens had been accumulated-- resulted in a food reward. If they chose the 'low risk' icon, only one token reward was given, but it was given whether the answer had been correct or not. Son and Kornell reasoned that if the monkeys knew whether or not they had made the correct response, that is, they had high confidence in their response, they should choose the high risk icon. If they were either not sure or knew they had made the wrong response, they should choose the low risk icon. And this is just what they did. The data showed that both monkeys were more likely to choose the high risk button when they had been right rather than wrong. The animals were also able to make

these confidence judgments appropriately about a dot density discrimination task. However, it might be possible to criticize these results on the grounds that the monkeys had simply learned to make a two-part response, through some shaping procedure, to a conditioned discrimination. The high confidence response might not have been analogous to a human confidence judgment about the choice, but instead might have been a shaped single response (which combined with the choice the way the multiple steps in the elephant putting its foot on the bucket, combine together) which was reinforced by the larger reward it drew.

Even more impressively, then, was the fact that these retrospective confidence judgments were observed to be appropriate immediately on a previously learned bone fide memory task--suggesting that they really were something like confidence judgments, rather than part of a single shaping sequence. Kornell, Son and Terrace (2007) showed transfer of the high risk/ low risk response on the first trial to a memory task that the monkeys had independently been trained to perform. The monkeys saw a series of 6 pictures and then had to do a recognition task in which they chose the correct picture from an array of one target and 8 distractors. After doing the immediate recognition task (and having the screen clear, so that the test alternatives, and their response were no longer in view), the monkeys were given the high risk/low risk icon choice. They immediately chose appropriately. The correlation between choosing high risk on trials on which they had given the correct response and low risk on trials in which they had not was significantly greater than zero for both monkeys. The three panels of figure 1 show Ebbinghaus first doing the memory task correctly, then being exposed to the confidence icons, and then expressing his high confidence in his correct choice.

Enter Figure 1 about here.

While the time lags in Hampton's (2001) and Kornell, Son and Terrace's (2007) tasks were both small and so the depth of the representation that was judged was not very impressive, they nevertheless were experiments in which the stimuli were not present in the environment when the judgment was made. In addition, in neither task were the test alternatives present when the judgment was being made. Furthermore, they were about memories; they were not conditioned discriminations. The rewarded stimulus changed on every trial in both experiments. These factors provide some reassurance that the animals may actually have been making some kind of assessments about their own knowledge--in the former case, whether they knew the answer or not, and in the latter, whether they had given the correct response or not. These experiments are the most rigorous that have given positive results suggesting that any nonhuman animal is capable of metacognition of any sort (even though the limitations on the metacognition are, of course, extreme). It appears that three monkeys alive today have metacognitive abilities.

Do any non-primates have metacognition?

We can, in good conscience, grant some limited metacognitive abilities to these three monkeys. Are any animals, other than primates, capable of metacognition? Of course, the answer must be that we do not know. Most animals have not been tested. However, Inman and Shettleworth (2001) and Sutton and Shettleworth (2007) have tested pigeons, and have concluded that they do not show evidence for metacognition. The task that the former used was somewhat similar to that used by Hampton (2003). It was a 3 -

alternative (rather than a 4-alternative) delayed-match-to-sample task. When the delay was increased, much as had been the case with the monkeys, the chance that the pigeons chose the escape (or uncertain) option increased. However, in striking contrast to the results found with the monkeys, who were able to do this task with above chance accuracy when the test stimuli were not present, the pigeons were unable to perform the task unless the test alternatives were present when they made their choice. This is telling. If metacognition entails a judgment about a memory or an internal representation, and the delay was needed to ensure that the judgment was about a representation, then this was the correct way to test for metacognition. The pigeons were unable to do it. And this is just what the researchers concluded.

Furthermore, Sutton and Shettleworth tried to elicit retrospective confidence judgments--similar to those studied by Kornell et al. (2007) from pigeons. Again, the birds were at chance unless the test stimuli were present. The conclusion , to date, is that although they have been tested, the results on pigeons indicate no metacognition.

Recently, Foote and Crystal (2007) have claimed, to much fanfare, that rats have metacognition. This conclusion, while well-publicized in the popular media, is far from universally accepted. Staddon, Jozefowiez and Cerutti (2007) have written a detailed rebuttal, based on risk assessment. Metcalfe and Terrace (submitted) also dispute Foote and Crystal's conclusion on these, and on other grounds.

Foote and Crystal (2007) trained 8 rats to do a duration discrimination task, in which a tone was heard for either a long time or a short time. The rats were given considerable training in this discrimination task, being reinforced for choosing the correct button to get a reward for 'saying' long--by choosing one button-- or 'saying' short--by

choosing the other button. In the next phase the rats were allowed to poke their noses into one hole if they 'wanted to take the test,' and into another hole if they did not want to take the test. If they chose to take the test, they were then given the button pressing test, and if they chose the 'long' button when the tone was long, they got 6 rat pellets. If they chose the 'short' button when the tone was short, they got 6 rat pellets. If, however, they chose the wrong button, they got nothing. A second hole for nose poking was introduced, and if they poked their noses into that hole--the 'don't take the test' hole --they got 3 rat pellets, regardless.

Rather than having only long and short durations, at the critical series of tests, the researchers included critical stimuli that were in between. Their logic was that if the trained up rats took the 'don't take the test' nose poke, selectively, when the stimuli were of intermediate length, then they would be indicating that they did not know. If they were more accurate when they decided to take the test than when they were forced to take the test, this, they thought, would be an indication of metacognition.

Data were presented data for 3 rats, who were more likely to choose the 'don't take the test' nose poke, when the stimuli were intermediate stimuli, than when they were either distinctively long or distinctively short. When those trials on which the animals were forced to take the test and those on which they chose to take the test were compared, they performed better with their own choice, on the difficult intermediate stimuli. These results were interpreted as indicating that the rats were metacognitive.

It is a clever experiment, and seems similar, on the surface, to that of Hampton, which did provide some evidence of metacognition. There are some critical differences, however. Most important is that the task was not a memory task, but rather a

conditioned discrimination task. It is not clear that mental representation or memory proper was involved, in this task, at all. The animals may simply have learned a three part discrimination. Second, there was no indication that the 'don't take the test' button meant that to the rats who chose it. Instead it may have been nothing more than a shaped multi-step response. There was no transfer test (such as Kornell et al. 2007) had used, to show that the meaning of the 'decline the test' button had any relevance to another task where the animal might also opt to decline the test.

How would a non-metacognitive animal do this task, to give the results obtained? As pointed out by Metcalfe and Terrace (submitted), the first thing to note is that only 3 out of the 8 animals did it. So, the first possibility is that it was simply accidental.

Second, the fact that there were two linked responses--the nose poke and the button press, can easily be explained by ordinary shaping behavior. The elephant rewarded for putting his foot on the bucket has first to put his other foot beside it. The initial nose poke may be no more than part of the complex rewarded pattern of motion that was reinforced over many trials. Finally, it is well known (from Pavlov on) that animals are responsive to intermediate categories in a conditioned discrimination task. Thus, the animals may well have been sensitive to the degree of discrepancy a test stimulus exhibited from the long and short stimuli on which it was trained.

What about the contingencies under the conditions in the experiment? The reward, in the case of a clear long or short tone was 6 pellets, as long as the animal got it right, which it nearly always did. If not the animal got 0 pellets. But the animal did not get the discrimination right when the stimuli were in the intermediate range. Indeed, the expected reward for tones exactly in the middle of the to-be- discriminated distribution

was 3. This was true if the rats 'decided to take the test,' in which case they had a 50-50 chance of being right, and getting 6 pellets or wrong and getting 0, yielding an expected gain of 3 pellets. It was also true if they decided not to take the test, in which case they got a sure 3 pellets. A non-metacognitive rat might have learned that if the to be discriminated stimulus was in the middle of the range, it did not matter what it did: the expected gain was 3 pellets, regardless. So it is not surprising to see that when the stimulus duration was extreme--either very long or very short-- the rats reliably do the thing they had been trained to do: poke their nose into the correct hole and choose the correct button. When the stimulus duration is in the middle --since it did not matter what the rat did, the expected gain is the same 3 pellets regardless--the rat is more likely to show random behavior. And that is exactly what the data show. No metacognition need be involved.

One more thing: why, on these intermediate stimuli, would the non-metacognitive rat be more likely to be right when it has poked its nose into the hole that the experimenters think meant that it wanted to take the test? The answer is simple. The stimuli in question had a correct answer, according to the experimenter's measurements--they were either slightly longer or slightly shorter in duration. They were not, in fact, exactly in the middle, where the odds were exactly the same for the different response combinations. When the rat perceived that a given stimulus was long (or short) it could get 6 pellets rather than 3. The difference in performance in the intermediate range of stimuli only indicates that the rats had some discrimination of stimulus duration, even in this range, and that the responses allowed them to use their own discrimination of the fine gradients when they were available. As Staddon et al. (2007) have noted, this variability

alone is enough to account for this seemingly convincing result. Rats, then, have not (yet) been shown to have metacognition.

Conclusion

Metacognition in humans provides them with the cognitive capability to assess their learning, their knowledge, and what would otherwise be their automatic responses to the stimuli in the world that drive behavior. How they do this has been the subject of intensive research (Blake, l973; Butterfield, Nelson, & Peck, l988; Costermans, Lories, & Ansay,1992; Dunlosky, Rawson, & Middleton, 2005; Hertzog & Dixon, 1994; Koriat, 1993; Sikstrom & Jonsson, 2005; Schneider, Vise, Lockl & Nelson, 2000). But not only do they have the capability to reflect on their mental representations, they take these reflections and put them to use in controlling how they will study (Metcalfe & Finn, in press; Finn, submitted), what they will choose to attempt to retrieve (Reder , l987; Reder & Ritter, l992), how they solve problems (Simon, 1979; Simon & Reed, l976), and how they will behave with respect to other people (Call & Tomasello, 1999; Wimmer & Perner, l983). All of these refined capabilities--both at the metacognitive and control level--are highly elaborated in humans. And although they are sometimes susceptible to biases and errors (Bjork, l994; Metcalfe, 1986), they nevertheless provide a buffer between what might correctly be called 'mindless' responding. Being reflections, that allow control of mental representations, these particular capabilities form the basis of what is usually referred to as 'mind' (Donald, l991; Suddendorf & Whitten, 2001). They are our escape from stimulus control and into self control.

Was Descartes right in attributing this kind of consciousness only to humans? Insofar as he was describing a highly elaborated self-reflective capability, the answer has

to be yes. However, that does not mean that Darwin (1859) was wrong. This capability, while highly developed in people shows antecedents in non-human species, most particularly in primates. To date, no studies with any animals other than primates has provided convincing evidence for this particular capability, though one has to be impressed by the remarkable non-metacognitive learning capabilities of non-primates, such as rats. Panskepp and Burgdorf (2003) for example, claims that rats laugh! There are a number of claims about the superior theory of mind capabilities of dogs. And perhaps most strikingly, the representational and time travel capabilities, as well as the deceptive capabilities, and episodic memory-like abilities of birds documented by Clayton (see, e. g., Dally, Emery, & Clayton, 2006) all seem astonishing. Perhaps, with further research, we will find traces of self-reflective consciousness--however elementary--in animals other than the three monkeys who have so far given evidence of some preliminary metacognitive capabilities.

References

Armstrong, D. (l968*). A Materialist Theory of the Mind.* London, Routledge and Kegan Paul.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. P. Shimamura, (Eds.), *Metacognition: Knowing About Knowing* (pp. 185-206). Cambridge, MA: MIT press.

Blake, M. (l973). Prediction of recognition when recall fails: Exploring the feeling-of-knowing phenomenon. *Journal of Verbal Learning and Verbal Behavior, 12*, 311-319.

Butterfield, E. C., Nelson, T. O., Peck, V. (1988). Developmental aspects of the feeling of knowing. *Developmental Psychology, 24,* 654-663.

Byrne, R. W., & Whiten, A. (1992). Cognitive evolution in primates: Evidence from tactical deception. *Man, 27*, 609-627.

Call , J. (2005). The self and other: A missing link in comparative social cognition. H.S. Terrace & J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness*. NY, NY: Oxford University Press.

Call, J., & Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition, 4*, 207-220.

Call, J., & Tomasello, T. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development, 70(2),* 381-395.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing

in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18,* 142 -150.

Dally, J. M., Emery, N. J., Clayton, N. S. (2006). Food-caching Western Scrub-Jays keep track of who was watching when. *Science, 312*, 1662-1665.

Damasio, A. (1994) *Descartes' error: Emotion, reason, and the human brain*, Putnam Publishing, 1994.

Darwin, C. (l859). On the origin of species by means of natural selection.

Descartes, R. (1637/1999). *Discourse on Method*. London: Penguin Books.

Donald, M. (1991). *Origins of the Modern Mind*. Cambridge, MA: Harvard University Press.

Dunlosky, J., Rawson, K. A., & Middleton, E. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*, 551-565.

Foote, A. L. & Crystal, J. D. (2007). Metacognition in the rat. *Current Biology, 17*, 551-555.

Frith, U., & Happe, F. (1999). Theory of Mind and self-consciousness: what is it like to be Autistic? *Mind & Language, 14,* 1-22.

Hampton R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences, 98(9),* 5359-5362.

Hampton, R. R. (2005). Can Rhesus monkeys discriminate between remembering and forgetting? . H.S. Terrace & J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness*. NY, NY: Oxford University Press.

Harrison, S. (2006). Augustine's way into the will, Augustine, Oxford Scholarship

Online Monographs.

Hertzog, C. & Dixon, R.A. (l994). Metacognitive development in adulthood and old age. In J. Metcalfe & A.P. Shimamura (Eds). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press (pp. 227-252).

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21(1)*, 101-1347

Hillyard, S.

Hochberg, J. (2003).

Humphrey, N. K. (1987). *The Uses of Consciousness*. New York: American Museum of Natural History.

Inman, A. & Shettleworth, S. J. (l999). Detecting metamemory in nonverbal subjects. *Journal of Experimental Psychology: Animal Behavior Processes, 25*, 389-395.

Jacoby, L. L., Bjork, R. A., & Kelley, C. M. (l994). Illusions of comprehensions and competence. In D. Druckman and R. A. Bjork (Eds.), *Enhancing Human Performance, III*. Washington, DC: National Academy Press.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609-639.

Leslie, A. M. (1987). Pretense and representation: Origins of "theory of mind." *Psychological Review, 94,* 412-426.

Loftus, E.F. (2004) Memories of things unseen. *Current Directions in Psychological Science*, 13, 145-147.

Metcalfe, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 623-634.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and a control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100,* 3-22.

Metcalfe, J. (l998). Cognitive Optimism: Self Deception or Memory-Based Processing Heuristics? *Personality and Social Psychological Review, 2*, 100-110.

Metcalfe, J & Terrace, H. (submitted) Rats still don;t have metacognition: A reply to Foote and Crystal.

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation*, vol. 26, (pp. 125-173). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition?. In J. Metcalfe, & A. P. Shimamura, (Eds.), *Metacognition: Knowing About Knowing* (pp. 1-25). Cambridge, MA: MIT press.

Oschner, K. N., &. Gross, J. J. (2006). The cognitive control of emotion. Trends in Cognitive Sciences, 9, 242, 250.

Panskepp, J. & Burgdorf, J. (2003). "Laughing" rats and the evolutionary antecedents of human joy? *Physiology and Behavior, 79*, 533-547.

Pavlov, I. P. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Translated by G. V. Anrep . London: Oxford University Press

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Povinelli, D. J. (2000). *Folk Physics for Apes*. NewYork, Oxford University Press.

Reder, L. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*, 90-138.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 435-451.

Rosenthal, D. (2002). Consciousness and higher-order thought, *Macmillan Encyclopedia of Cognitive Science*, Macmillan Publishers Ltd, pp. 717-726

Russell, B. (1945/1972). *A History of Western Philosophy*. New York: Simon & Schuster.

Shields, W. E., Smith, J. D., & Washburn, D. A. (1997). Uncertain responses by humans and Rhesus monkeys (Macaca mulatta) in a psychophysical same-different task. *Journal of Experimental Psychology: General, 126,* 147-164.

Shields, W. E., Smith, J. D., & Washburn, D. A. (2005, BBS

Sikström, S. & Jönsson, F. (2005). A model for stochastic drift in memory strength to account for judgments of learning. *Psychological Review, 112*, 932-950.

Simon, H. A. (1979). Information processing models of cognition. *Annual Review of Psychology, 30,* 363-396.

Simon, H. A., & Reed, S. K. (1976). Modeling strategy shifts in a problem-solving task. *Cognitive Psychology, 8(1),* 86-97.

Simons, D. J., & Chabris, C. F. (l999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*, 1059-1074.

Schneider, W., Vise, M., Lockl, K, & Nelson, T. O. (2000) Developmental trends in children's memory monitoring : Evidence from a judgment of learning task.

Cognitive Development, 15, 115-134.

Son, L. K. & Kornell, N. (2005). Meta-confidence judgments in Rhesus Macaques: Explicit versus implicit mechanisms . H.S. Terrace & J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness*. NY, NY: Oxford University Press.

Terrace, H. S., & Metcalfe, J. (2005). Introduction. In H.S. Terrace & J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness*. NY, NY: Oxford University Press.

Treisman, A.

Staddon, J. E. R., Jozefowiez, J.,& Cerutti, D. (2007) Metacognition: A problem not a process :"Metacognition" in animals can be explained by familiar learning principles, *PsyCrit*

Suddendorf, T. Whiten, A. (2001). Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological Bulletin, 127(5),* 629-650

Sutton, J. E., & Shettleworth, S. J. (2007). Pigeons still don't have metamemory. Paper presented at the Annual Meeting of the Comparative Cognition Society.

Terrace, H. S. (2005). Metacognition and the evolution of language, In H.S. Terrace & J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness*. NY, NY: Oxford University Press.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. New York, NY: The Century Co.

Tomasello, M., & Call, J. (1997). *Primate Cognition*. New York: Oxford University

Press.

de Waal F. B. M. (1992). Intentional Deception in Primates. *Evolutionary Anthropology, 1,* 86-92.

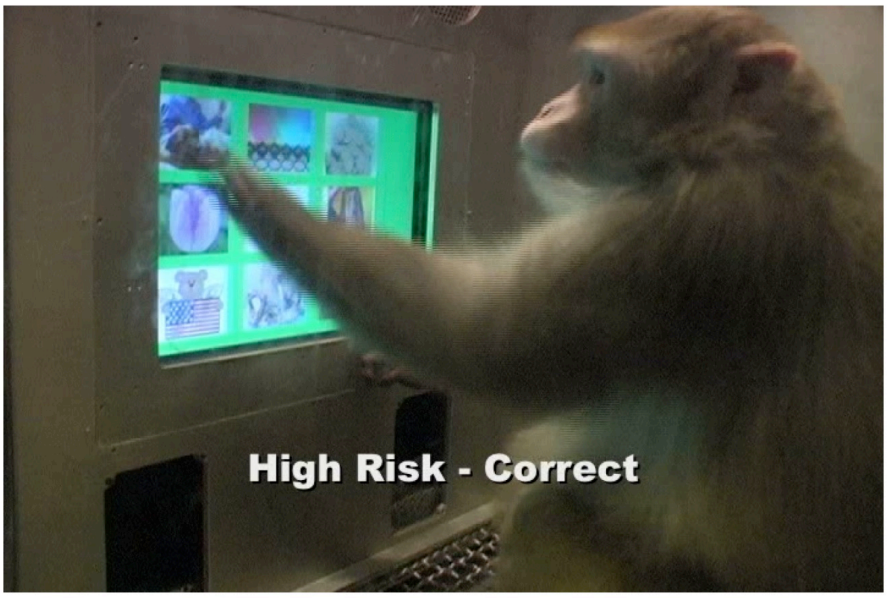
Wager, T. D. & Smith, E. E. (2003). Neuroimaging studies of working memory: A meta-analysis, Cognitive, Affective and Behavioral Neuroscience, 3, 255-274.

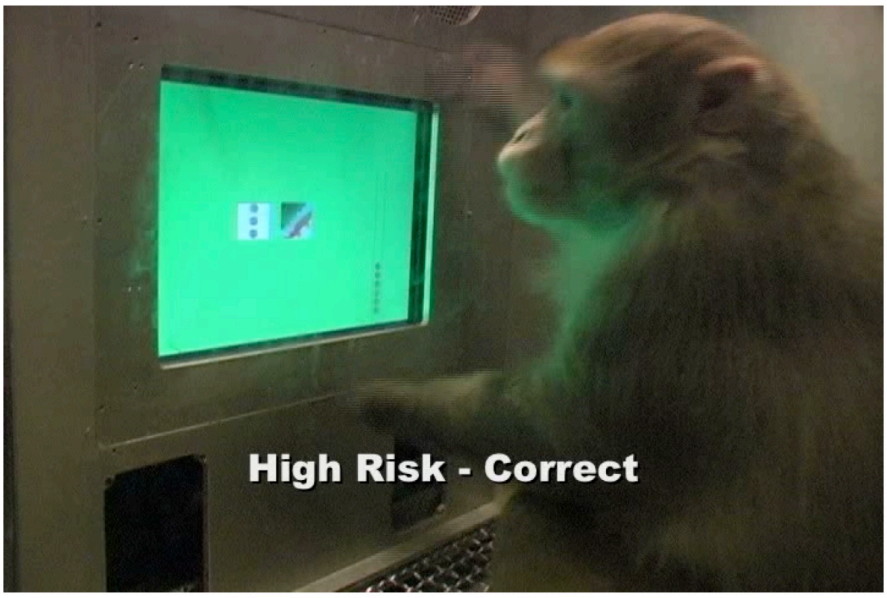
Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates. *Behavioral and Brain Sciences, 11, 233*-273.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103-128.

Figure caption

Figure 1 a,b,and c. Panel a shows Ebbinghaus correctly choosing the to-be-remembered item in a recognition task. Panel B shows him thinking when the confidence icons appear. Panel C shows him choosing the high risk (high confidence) icon.


High Risk - Correct


High Risk - Correct


High Risk - Correct