Comparative metacognition
Herbert S Terrace¹ and Lisa K Son²

Metacognition is knowledge about knowledge, often expressed as confidence judgments about what we know. Most of the literature on metacognition in humans is based on subjects’ verbal reports. Investigators of animal cognition have recently described nonverbal methods for investigating metacognition in animals. In one, subjects are given the option to escape from difficult trials. In another, subjects are trained to place bets about the accuracy of their most recent response. To rule out noncognitive interpretations of purported evidence of metacognition in animals, one must ensure that escape responses do not increase the overall density of reinforcement and that they do not occur in the presence of the stimuli on which the subject was trained. The nonverbal techniques used to investigate metacognition in animals make possible two interesting lines of research: investigating the contribution of language and explicit instruction in establishing metacognition, and the investigation of the neural substrates of metacognition.

Addresses
¹ Columbia University and the New York State Psychiatric Institute, United States
² Barnard College, United States

Corresponding author: Terrace, Herbert S (terrace@columbia.edu) and Son, Lisa K (lson@barnard.edu)

“Know thyself”, attributed to the Delphic oracle, Socrates and Solon.

It can be safely said that the Delphi oracle did not have animals in mind when she uttered her famous dictum about self-knowledge. That possibility was not even considered until the latter part of the 20th century, when evidence began to accrue that animals have minds. The evidence came from experiments in which behavior could not be explained by reference to stimulus–response associations between observable events because the behavior in question occurred in the absence of the stimulus to which the subject was trained to respond [1]. Accordingly, the control of those behaviors was attributed to a representation of the training stimulus and the animal’s mind was considered to be the repository of such representations [2].

Metacognition is arguably the most complex form of cognition studied in animals because it requires the animal to form a representation about a representation (as opposed to a single representation, as in other examples of animal cognition) [4,5]. But the complexity of metacognition in animals is miniscule when compared to the complexity of human metacognition as observed in experiments on thinking, learning, and problem solving [6,7].

Experiments on metacognition in animals raise a slew of intriguing questions. How similar is metacognition in animals, which cannot be based on verbal knowledge, to human metacognition? Is language necessary for human metacognition? Are the neural mechanisms that give rise to human and animal cognition analogous? At present, the answers to these questions are sketchy at best. At worst, they are inchoate in the sense that comparisons of models of human and animal metacognition suffer from the acute absence of relevant data.

The most direct point of comparison between studies of human and animal metacognition is at the level of prospective and retrospective metacognition. Retrospective metacognition occurs when people make confidence judgments about their knowledge of prior events; prospective metacognition, judgments about future events, for example, confidence in one’s performance on an upcoming test. In a typical experiment on prospective metacognition, subjects are asked to engage in some cognitive task, say, memorizing a list of paired-associates. They are then asked to rate their confidence about responding accurately on a test on which they would have to recall which cues were paired with which associates. Once the test is administered, a correlation between subjects’ accuracy on the test and their pretest confidence is used as a measure of subject’s metacognitive knowledge [8]. Similarly, to assess retrospective metacognition, subjects would be asked to make judgments about their confidence in the accuracy of their responses after taking a test.

Although people are fairly good at judging the accuracy of their own knowledge, there has been a consistent tendency to be overconfident [9–11]. Such faulty metacognition was an important instigator of a surge of
research on cognitive mechanisms underlying metacognitive judgments [12]. Several theories have been proposed to investigate this question, but none are able to explain a majority of the existing data [13–15]. One explanation, known as target retrieveability or accessibility, is perhaps the most intuitive [16]. It postulates that people base their judgments on the number of features of the target they can access or retrieve that is the more features that are retrieved, the higher their confidence. By contrast, the inferential view states that people base their judgments on the basis of cue-familiarity rather than target-familiarity [17].

**Metacognition in animals**

Table 1 provides a summary of recently developed non-verbal paradigms to assess metacognition in animals. Each paradigm assumes that subjects experience varying degrees of uncertainty that is inversely related to the difficulty of the task. Because verbal reports are typically used to assess metacognitive knowledge, it has not been possible to determine the extent to which metacognition depends on language, and whether this phenomenon is uniquely human. Although we can be sure that metacognition in animals does not require verbal ability, it is unclear to what extent metacognition in animals is analogous to that observed in humans. To see why, we have assigned recent experiments on animal metacognition shown in Table 1 into one of four categories. The main contrasts to be drawn in Table 1 are whether or not an animal’s purported metacognitive judgments were made in the presence of the stimuli on which they were trained (Category 1), whether that judgment was differentially reinforced (Category 2) and whether the judgments they made were prospective (Category 3) or retrospective (Category 4).

**Category 1**

The earliest studies of animal uncertainty make up the largest category [18–29,30*,31]. Unfortunately, the results of these studies are ambiguous because it is possible to account for purported examples of metacognitive behavior with simpler noncognitive explanations. In Table 1, we refer to such judgments as ‘concurrent’ because the purported metacognitive response is made in the presence of the stimuli the subject is asked to discriminate. Consider, for example, Smith _et al._’s influential study in which dolphins were required to discriminate the auditory frequencies of two tones by responding to one of two stimuli [18]. Whenever the tone was exactly 2100 Hz, a response to a ‘<2100 Hz’-icon’ was rewarded; for lower frequencies, a response to a ‘<2100 Hz’-icon’ was
showed that ability when trained on a new task [35]. Previous experiments, it is not surprising that they had extensive metacognitive training in their only option. Accuracy on catch trials was reliably lower than it was on trials on which subjects could choose the test option. That difference shows that subjects could anticipate the outcome of a trial.

**Category 4**

The contingency for reward in the experiment listed in Category 4 was based on a subject’s ability to remember the accuracy of its response on a particular trial [35**]. If, for example, a monkey was trained on a serial probe recognition (SPR) task, reward during baseline training was determined solely by the subject’s ability to recognize an item that was presented during the sample. During a test for metacognition, reward was based solely on its confidence it had about the accuracy of its response on the SPR task.

The low-confidence and the high-confidence icons were presented following a subject’s response on the SPR task. As shown in Figure 1B, reward and punishment consisted of, respectively, the addition and the subtraction of tokens that were displayed in a bank on the right side of the subject’s monitor. Choosing the high-risk icon following a correct response on the SPR task resulted in the addition of three tokens. Choosing the high-risk icon after an incorrect response resulted in the removal of three tokens. In order to maintain responses to the low-risk icon, correct and incorrect responses to that icon were each rewarded by the addition of one token; however, the subject responded.

Subjects’ choices of risk icon were highly accurate. Other features of their performance provided new evidence of...

---

8 Because responses to the third stimulus could increase the overall frequency of reward, it would be helpful if the authors of articles that use this procedure estimate how many ‘extra’ rewards a subject would have lost if that contingency were absent.

9 In the SPR task, a multi-image sample containing n items, is displayed successively. During test, the subject is presented with a single item from the sample. The subject’s task was to choose in item that was included in the sample.
retrospective metacognition in animals. As shown in Figure 2, there was immediate and complete positive transfer from metacognitive tests that were administered following training on one task to a qualitatively different task, for example, from tasks on which subjects were trained on various types of magnitude discrimination tasks to an SPR task that was a test of short-term memory [36]. Positive transfer of metacognitive ability across qualitatively different tasks suggests that it was readily available in monkey’s cognitive tool kit.

Another type of evidence was derived from an analysis of the contribution of reaction time (RT) to metacognitive judgments. In experiments on human metacognition, RTs are typically shorter for correct than for incorrect responses [37]. A similar difference was observed in the case of monkeys, that is, a negative correlation between RTs during test and the accuracy of their choice of risk icon. Although there is no simple method to control for the influence of RT duration on confidence judgments, it is possible to partial out that factor. The resulting values were significantly greater than zero. It follows that the duration of RTs on correct and incorrect trials could not be the sole determinant of subjects’ metacognitive performance.

Like RTs, other peripheral or external cues may influence metacognition, for example, item difficulty. Figure 3 presents the percentage of trials on which a ‘high-confidence’ bet was made at each level of difficulty (top two panels) and the SPR task (bottom two panels). Level of difficulty was measured by the size difference between the target and distractors on a circle-size test and by serial position on the memory test. As can be seen, subjects could have used difficulty as a cue for making their metacognitive judgments, for example, ‘choose high risk when the difference between target and distractors is sufficiently large’. However, the differences in level of difficulty cannot explain the risk selections on the SPR task because the relative frequency of accurate high-confidence bets was essentially the same at virtually every serial position. In other words, monkeys assessed the
Monkeys can not only make reliable judgments about the accuracy of their responses, while ignoring obvious external cues like serial position.

A variation of the paradigm on which the rats were originally trained addressed that problem by obtaining behavioral measures of the subjects’ uncertainty, similar to those obtained from experiments in Category 1 of Table 1. Instead of simply waiting for the outcome of a trial, subjects were free to terminate it by withdrawing from port 1 or 2. They could start a new trial by entering a third port that was equidistant from ports 1 and 2. Subjects’ behavior matched that observed in similar experiments in which subjects were allowed to abort a trial and to start a new trial immediately. Because the rate at which subjects opted to start a new trial varied with the difficulty of the discrimination on a particular trial, Kepecs et al. concluded that ‘a key function of OFC is to generate reward predictions based on stimulus-reward associations...and that OFC neurons signal outcome prediction.’ However, like the other experiments listed in Category 1, the results of Kepecs et al.’s experiments are ambiguous because the same results could have been obtained if the subjects followed the simpler strategy of maximizing reinforcement or if they responded to the small difference between the discriminative stimuli as a cue for making an escape response. It would therefore be of interest to replicate the Kepecs et al. experiment using a paradigm in which subjects expressed their confidence about the outcome of a particular trial after the odors were removed from the cones, and in which responses to the purported indicators of confidence did not influence the rate of reward.
Explicit and implicit metacognition in humans

Given that a nonverbal task such as the betting paradigm provides evidence that animals are able to express uncertainty, it may be possible to pursue other interesting questions about metacognition in animals that parallel those that have been investigated with human subjects. One important issue is, what is the relationship between explicit knowledge and metacognition \[41,42\]? While some would agree that explicit knowledge improves metacognitive accuracy \[43\], others would even argue that ‘to make conscious a metacognitive process, is interfering with the task at hand.’ \(44\), p. 70).

While verbal subjective reports may prove valid in a typical laboratory setting, their ecological validity might be much stronger if actual bets were made and carried out as they would be in the real world. A recent study of young children, ages 5–6, used the nonverbal betting paradigm, in conjunction with an SPR task, because firstly, it was expected to be easier to comprehend than the typical 10-point rating scale of confidence; secondly, it might increase the ecological validity of metacognitive judgments; and thirdly, it would provide a comparison between the performance of humans and monkeys on the same metacognitive task \[45\]. In one condition, no verbal instructions were given about how to bet, which is the method used with monkeys. In another, the children were given explicit instructions on how to bet. Results showed that both groups learned to bet appropriately that is the correlations between accuracy on the SPR test and the size of a bet were positive, although, as expected, higher for the explicit group. More importantly, when given a transfer test on which subjects were asked to respond to multiple choice vocabulary questions, children trained by

---

**Figure 3**

The proportion of trials on which two monkeys selected ‘high risk’, for varying levels of difficulty on the circle-size task (top two panels) and the SPR task (bottom two panels). On the circle-size task, the monkeys had to choose either the smallest or largest circle (depending on monkey) from an array of nine circles. Difficulty was determined by the difference in size from the target and the distractors (which were all the same size). On the SPR task, monkeys had to judge whether a probe stimulus that was presented after a sample containing six successively presented items was included in the sample or not (familiarity task). Difficulty was determined by the serial position of the target.
the explicit condition performed better than children trained by the implicit ‘monkey’ condition (although both groups exhibited positive correlations). These results suggest that, while metacognitive processes may not require explicit awareness, declaring what you know during training is beneficial.

Conclusion

Metacognition has developed as an important field of inquiry in both human and animal cognition, but much more research is needed in each instance to provide a basis for direct comparison. In the case of humans, there are enough data to test specific models of metacognition [46,47], but none is available in the case of animal metacognition. It is also important to investigate experiments on the development of metacognition in young children to determine the extent to which metacognition requires language. In the case of animals, there is firm evidence that monkeys can make both prospective and retrospective confidence judgments about their performance on various recognition tasks, in the absence of language, but much more research is needed to define the relevant parameters. Two obvious advantages of studying metacognition in animals are that it makes it possible to study nonverbal metacognition and to investigate the neurological underpinnings of such judgments. However, before that can happen, agreement is needed as to what constitutes a valid marker of animal metacognition. Specifically, noncognitive factors have to be eliminated from purported demonstrations of metacognition in animals. Another problem when comparing animal with human metacognition is that monkeys have had far less training than human subjects at the start of an experiment. Other experiments that could clarify possible differences in the metacognitive abilities of animals and humans should address firstly, the extent to which a monkey could tolerate intervals between the response on a cognitive task beyond 1 s, the current upper limit and secondly, a monkey’s ability to make finer distinctions of its confidence, than ‘confident’ and ‘not confident’ when making metacognitive judgments. The extent to which an animal can maintain or recall its metacognitive knowledge and make fine distinctions between different levels of confidence should help clarify its relationship to human metacognition.

Acknowledgements

We would like to acknowledge support for preparing this article from a grant from NIMH: R01 MH081153-01A1 and a grant from the DOE, Institute for Education Sciences.


Metacognition has always been thought to play a causal role: if people know what they know and do not know, they can then choose effective learning strategies, which will in turn benefit later performance. Similarly, if any pieces of this causal role breakdown (e.g., if people cease studying prematurely) then later performance will be less than optimal. In this recent study, evidence for the causal role of metacognition was found.


A crucial question in the field has been to understand how people’s metacognitive knowledge guides study behavior. The most common method that has been used to address this issue has been to observe how much study time learners allocate to a variety of materials. The majority of early data have shown that people spend most of the time on items that were very difficult, supporting a ‘set-threshold’ hypothesis. However, since the early 1980s, more research is needed in each instance to provide a basis for direct comparison. In the case of humans, there are many research methods that are available to study this issue; however, in the case of nonhuman animals, there are far fewer methods available to study this issue, and no model is available in the case of animal metacognition.


This article describes a series of elegant experiments that were the first to investigate the neural basis of performance on a metacognitive task. It distinguished the firing of single cells in the frontal/orbital cortex on the basis of their responses to the difficulty of an olfactory discrimination, whether responses were correct or errors and subject's confidence that it could respond correctly on the current trial and thereby earn a reward and avoid a lengthy TO. The latter distinction suffered from the same methodological problem that was made in many behavioral studies of metacognition. In this instance, there is no way to distinguish between two interpretations: firstly, that the behavioral responses in question were based on a subject's metacognitive judgment of low confidence that it could respond correctly on a particular trial or secondly, the maximization of reward.


This was the first study of metacognition in animals that was based on the memory of previously presented stimuli, rather than on the perception of stimuli that were present at the time at which a subject responded to a high-risk or low-risk confidence icon. It was also the first experiment based on prospective judgments. A metacognitive task that is based on memory rules out the possibility that the purported metacognitive response was under the control of an external stimulus. If it were, it would be indistinguishable from other discriminative responses for which it is not necessary to postulate a metacognitive judgment.


This was the first study of metacognition in animals that was based on the memory of stimuli and the responses made to those stimuli (retrospective judgments). It was also the first experiment to use a transfer test to show that subjects were able to make metacognitive judgments on a task that was qualitatively different from the one on which they were trained. In this instance, the transfer was from a magnitude discrimination task to a short-term memory task. Another novel feature of this study was the use of a hint paradigm in which subjects learned to seek hints on a need-to-know basis. Hints were requested when subjects were making errors during initial training but not when subjects learned the task.


