12

Metaconfidence Judgments in Rhesus Macaques: Explicit Versus Implicit Mechanisms

Lisa K. Son & Nate Kornell

To date, psychological research on human metacognition centers on whether and to what extent people know what they know. For example, feeling-of-knowing judgments ask people how certain they are that they know an answer that they are unable to retrieve. Judgments of learning ask how certain they are that they will be able to remember a recently learned item in the future. Confidence judgments ask people how certain they are that their response to a question is correct. In this chapter, we focus on confidence judgments, which in humans are typically made verbally. Our chapter describes an experiment on confidence judgments in rhesus macaque monkeys. Because of their obvious lack of verbal ability, the monkeys were given the opportunity to express their confidence by placing bets on the accuracy of their responses in a cognitive task.

BACKGROUND ON HUMAN METACOGNITION

Until recently, the experimental study of metacognition had generally been limited to the human species. We begin by summarizing the major theories of human metacognition before discussing their extension to nonhuman species. The term "metacognition" has been associated with uniquely human qualities such as introspection, self-reflection, frontal lobe function, theory of mind, and even consciousness (Flavell, 2000; James, 1890; Janowsky, Shimamura, & Squire, 1989; Shimamura & Squire, 1986; Tulving, 1994; Tulving & Madigan, 1990). For example, Tulving (1994) says that metacognitive researchers use "behavioristically safe expressions, such as memory 'monitoring,' mnemonic 'behavior,' memory 'search,' tip-of-the-tongue 'states,' and

feeling of knowing 'experience' . . . possibly to avoid the big bad 'C' word" (p. ix), "C" being consciousness. Metcalfe and Kober (chapter 2, this volume) say that metacognition is associated with having a self-reflective "inner eye" that can look at other cognitive functions and content. Thus, the ability to self-reflect, or introspect on one's internal mental representations, has been considered a foundation for human consciousness (see Metcalfe & Shimamura, 1994, preface).

Metacognitive processes, which have been more formally specified by Nelson and Narens (1990, 1994), consist of a basic structure containing two interrelated levels, an object-level and a meta-level. The two levels are shown in figure 12.1A. The object-level may include an

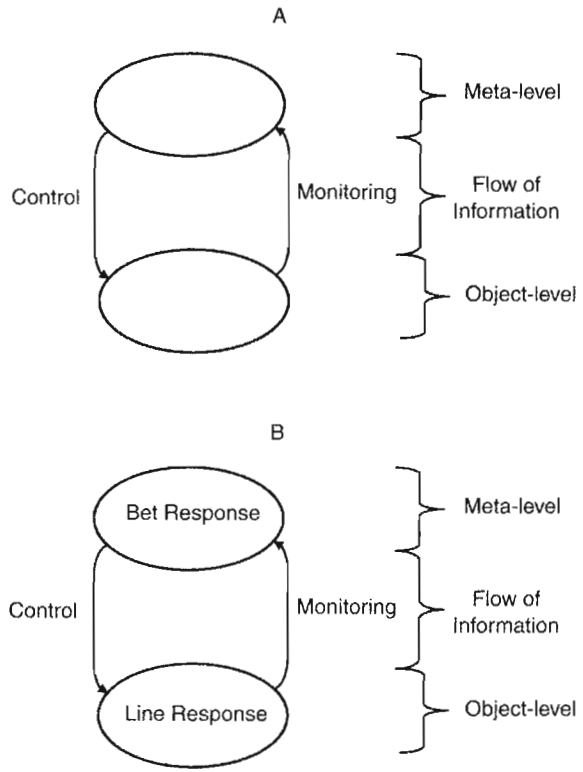

Figure 12.1 A. An overview of the two separate, but interacting, processes during learning as illustrated by Nelson and Narens (1990, 1994). During metacognitive monitoring, the meta-level is informed by the object-level of the present state, and, in turn, during metacognitive control, the meta-level modifies the object-level. B. The current investigation: A meta-level confidence judgment, as measured by risk response, assessing an object-level line task.

individual's memories, cognitions, and behaviors, and describes the state of the present situation. The meta-level monitors (and sometimes can control) the object-level. The interaction between the two levels has generally been thought of as being one of self-reflection. Nelson and Narens (1994) described people as "systems containing self-reflective mechanisms for evaluating (and re-evaluating) their progress and for changing their ongoing processing" (p. 7). The results of processing at the meta-level are judgments and feelings, such as confidence judgments and other expressions of certainty.

One important line of research within the area of metacognition is the issue of how the metacognitive judgments, or meta-judgments, are made. Two categories of mechanisms have been proposed to underlie meta-level judgments in the human literature: direct-access mechanisms and inferential mechanisms (see Nelson, Gerler, & Narens, 1984; Schwartz, 1994). Direct-access mechanisms rely on the accessibility of the internal memory trace. An example is the metacognitive judgment of being in a tip-of-the-tongue state (TOT)—in a TOT, people can often retrieve part of the target memory (e.g., only the first letter), causing a state of frustration. But it is the state of being "close" to accessing the internal memory that causes the high feeling of knowing in a TOT, and the more features of the memory one is able to retrieve, the stronger the meta-judgment of certainty.

Some of the earliest studies of metacognition were interpreted as supporting the direct-access view. For example, in the first investigation of feeling-of-knowing judgments (FKJs), Hart (1965) asked participants trivia questions, and then asked for FKJs on those items the participants could not answer. Then a final recognition test was given. Results showed that people's judgments were positively correlated with recognition on the final test—items given high judgments were more likely to be recognized than items given low judgments. The results indicated that people's judgments were probably based on an accurate assessment of what was in memory. Underwood (1966) conducted a similar study, presenting participants with three-letter trigrams. The trigrams varied from common three-letter words to difficult consonant syllables. Participants had to judge the difficulty of learning the items, after which they were all given a recall test. Results showed that individuals predicted their own recall with high accuracy. Also, Arbuckle and Cuddy (1969) presented participants with lists of paired associates, and asked them to predict whether they would be able to recall each pair. After giving the judgment, the participants were given a memory test. Results showed that the predictions were accurate—those paired associates given lower ratings were recalled less well than those given higher ratings. The main conclusion from these early studies was that people's meta-judgments were based on the amount of information they had been able to retrieve, or directly access, from an internal memory representation at the time of judgment.

Unlike the direct-access view, the inferential view posits that meta-judgments are based on cues that are external, rather than internal, to the memory trace. For example, a judgment may be made based on how familiar the topic of the question is (e.g., giving a high judgment to the question, "Who is the tallest basketball player in the NBA?" because basketball is a familiar topic). Or a judgment may be based on how quickly the answer comes to mind (e.g., giving a higher judgment following a quickly retrieved answer than a slowly retrieved answer). Both of these can predict the accuracy of a memory, but neither relies on directly accessing information about the internal representation.

Schwartz and Metcalfe (1992) compared the direct-access and inferential accounts in an experiment by comparing the relative contribution of the memory strength of the cue and target in making FKJs. They presented participants with paired associates and asked for FKJs in presence of only the cue. Prior to presenting the list, they also manipulated the accessibility of some of the cues by preexposing them in a pleasantness-rating task, and they did the same for a different set of target items. Their results showed that priming cues influenced FKJs, but priming targets did not. They concluded that the strength of the memory trace did not influence meta-judgments, as a direct-access account would predict; instead the judgment was inferred from the familiarity of the cue (see also Glenberg, Sanocki, Epstein, & Morris, 1987; Metcalfe, 1993a, 1993b; Metcalfe, Schwartz, & Joaquim, 1993; Miner & Reder, 1994; Reder, 1987; Reder & Ritter, 1992).

Benjamin, Bjork, and Schwartz (1998) found further evidence for inferentially based meta-judgments. They had participants answer general information questions, such as "What is the color of topaz?" Participants were also told that the time it took them to answer each question was of primary interest, and to press the Enter key as soon as they knew the answer to the question. After answering each question, participants gave a judgment predicting how well they would be able to remember their answer on a later test on a scale from 0 (no chance of later recall) to 100 (certain later recall). The data showed what was called a retrieval fluency effect—participants' judgments were negatively correlated with response times for answering the questions. The researchers concluded that the speed or ease with which one is able to come up with an answer seems to be a source of information for making meta-judgments—even though, in this case, it led to very inaccurate judgments. Furthermore, the judgments could not have been based on the amount of information directly accessed from the memory trace, because all of the answers were fully retrieved. Based on these studies, then, there is ample evidence that meta-judgments can

be based on inferential or external cues—cues other than those based on an assessment of an internal representation.

The results of these studies point out that meta-judgments need not be based on introspection of the internal memory trace. Instead, a host of inferential or external cues may be the basis of our judgments. An interesting question, then, might be to ask whether awareness of such cues is necessary in order to make the judgment—some think not (see Cary & Reder, 2002, an article titled, "Metacognition . . . Giving Consciousness Too Much Credit"). Below, we present data that suggest that meta-judgments need not be based on decisions that one is aware of, but instead may be based on implicit mechanisms.

IMPLICIT METACOGNITION

Outside of psychology laboratories, people are not asked to verbally provide meta-judgments very often. However, we believe that they may, nevertheless, be made frequently and, more important, without our awareness. To illustrate, most people, if asked what Lyndon Johnson had for dinner on February 23, 1958, would probably feel uncertain and say, "I don't know." When asked their own name, they would feel very certain and would state their name immediately. It seems rather absurd that meta-judgments such as "I'm quite sure that I know my name" would or should reach awareness. Yet, merely responding "I don't know" might contain a tacit meta-judgment of uncertainty. Here is another example: On the game show *Jeopardy*, a successful contestant only rings her buzzer when she feels fairly certain that her answer is correct (or in *Jeopardy* terms, she feels fairly sure that her question is correct). Her decisions to press the buzzer must be made very quickly, perhaps even more quickly than the time needed to become aware of her decisions. Still, each of those decisions may consist of several different tacit meta-judgments.

In several experiments using a game-show paradigm similar to *Jeopardy*, Reder and colleagues (e.g., Reder, 1996) asked participants a series of trivia questions. They were told to imagine that they were competing against another contestant, and to say as quickly as possible whether they knew the answer to the question. Only then would they be able to answer the question to earn points. Results showed that these meta-judgments could be made more quickly than participants could retrieve the answers. Furthermore, the judgments were usually accurate in predicting subsequent accuracy of the answer. Reder (1996) also found that priming words in the question led to increased subjective estimates of knowing the answer, despite the fact that this exposure did not improve actual rates of producing the correct answer. Reder and her colleagues concluded that some meta-judgments operate at an implicit level (Reder, 1996; Reder & Schunn, 1996). We believe that certainty and uncertainty are feelings that can occur to us even when we cannot explain them and can result from implicit cues that are always present. For the most part, it is only when we are asked about our level of certainty that we become aware of that certainty. As we will see later in the chapter, the implications of animals making judgments of uncertainty depend very much on whether one views the judgments as being an implicit, nonverbal process or a conscious experience.

In this chapter, we relate explicit mechanisms with that which pertains to an assessment of an internal representation, mainly because this requires that one directly access retrieved pieces of information from the memory trace. The notion that meta-judgments are based on explicit mechanisms fits nicely with the original, more philosophical definition of metacognition of labeling humans as self-reflective machines. On the other hand, we relate implicit mechanisms with meta-judgments made on the basis of external cues that are currently present, mainly because these judgments may be made without any internal monitor. These judgments transform continuously as a result of the constantly changing external cues and transient events that people are typically unaware of. Based on the data in the human literature summarized above, we would argue that people use both explicit and implicit metacognitive processes. The question of whether any metacognitive processes exist in nonhuman species remains.

ARE METACOGNITIONS SPECIAL TO HUMANS?

As mentioned earlier, metacognition is considered a very high-level mental function unique to humans. "Machines without consciousness, and animals whose consciousness is different from that of human beings, could not perform many of the tasks that human subjects in metacognitive experiments, and others of the same general kind, can and do perform" (Tulving, 1994, p. ix). Tulving and Madigan (1970) state that metacognition is "one of the truly unique characteristics of human memory: knowing about knowing" (p. 477). Metcalfe and Shimamura (1994) begin their book by saying, "The ability to reflect upon our thoughts and behaviors is taken, by some, to be at the core of what makes us distinctively human" (p. xi).

One obvious difference between humans and animals is the ability to speak. A question, then, is whether it is the ability to verbalize thoughts and feelings that grants humans the ability of self-reflection. And if one did not—or more pertinently, could not—verbally express one's feelings, does that then rule out the possibility that meta-judgments can occur? We think not. Thoughts or feelings, after all, are not synonymous with verbal speech. Judgments that are based on the familiarity of a question or cue, for instance, do not seem too cogni-

tively advanced for an animal. After all, animals can easily discriminate more familiar stimuli from less familiar ones. In addition, anyone who has seen a dog hesitate before jumping into a high truck bed, or a monkey waver between two choices in a psychological task, is likely to imagine that the animal is experiencing some sort of uncertainty. Of course, there is a difference between behaving in a way that appears uncertain and being able to report feelings of uncertainty about cognitions. The latter is analogous to answering the question, "Can you give me a judgment about the certainty of your response?" and only it qualifies as metacognitive.

A few empirical studies have shown that monkeys and dolphins are able to make uncertainty responses (Hampton, 2001; Shields, Smith, & Washburn, 1997; Smith et al., 1995; Smith, Shields, Allendoerfer, & Washburn, 1998; Smith, Shields, Schull, & Washburn, 1997; also see chapters 10 and 11, this volume). In an early study of uncertainty in animals conducted by Smith et al. (1995), human and dolphin subjects were asked to respond *high* when a 2100-Hz tone was played, and *low* for tones less than 2100 Hz. Then they were tested with tones ranging from 1200 Hz to 2100 Hz, with the correct response being *low* for any tone but the 2100 Hz tone. However, in addition to the *high* and *low* choices, a third choice, *escape*, was offered simultaneously, which, when pressed, returned a guaranteed reward, which was smaller than the reward for a correct *high* or *low* answer. The results showed that both humans and dolphins responded *high* for 2100-Hz tones and *low* for low tones up to about 2080 Hz. Most interestingly, they chose *escape* on trials with tones in between. Presumably these were the most difficult trials, where the subjects would have felt the most uncertain about their answers. Follow-up studies showed similar evidence of uncertainty in monkeys.

No one would question that human participants chose to escape difficult trials because they felt uncertain. But is this conviction based on the fact that people can verbalize the feelings by saying, "I chose to escape because I was uncertain"? Or is the behavior (pressing the escape response) enough for us to assume feelings of uncertainty in human participants? If so, then we could grant those same abilities to animals that behave similarly. However, when animals do behave similarly, the immediate reaction often is to ask whether the results might be based on external cues, or whether the learning was based on simple stimulus-response associations, or whether these behaviors did not require any internal monitoring at all. For example, could the animals have simply learned that to maximize reward, high tones should be followed by *high*, low tones should be followed by *low*, and middle-level tones should be followed by the escape response? In this case, no meta-level process is necessary; object-level processes suffice.

Hampton's procedure (see Hampton, 2001; chapter 11, this volume) investigated prospective judgments in monkeys by using a modified escape procedure. He presented two monkeys with a picture. Following a variable delay, the monkeys were given a forced recognition test in which they had to identify the picture that they had been shown among three distractors. However, on some of the trials, rather than being forced to take the test, the monkeys were free to either choose to take the test (for a big reward only if correct) or escape the test (for a guaranteed smaller reward). If the monkey chose to escape, the trial ended without the test being taken. This decision was made prospectively, before the test was presented. His results showed that both monkeys performed better on tests that they freely chose to take than on tests that they were forced to take, suggesting that the monkeys chose to take the test when they were certain. The improvement in this procedure over the original escape procedure is that the test was not present simultaneously with the escape choice. However, a few issues remain. For example, when the monkey chooses to escape the trial, the trial ends, and no object-level response is made. Thus, on a given trial, there is no way of knowing the monkey's actual performance on the test and his certainty of taking the test. In the rest of this chapter, we present findings using a new paradigm in which meta-judgments can be measured separately from object-level responses in nonhuman animals.

META-CONFIDENCE JUDGMENTS IN ANIMALS

Is there a way to measure meta-level judgments, as in Nelson and Narens' (1990) framework in animals? We decided to investigate this question using confidence judgments made about previous responses. Using confidence judgments seemed like the obvious next step, particularly since others, mentioned above, had already shown that animals could report uncertainty by using the escape response efficiently. In our task, two responses were required, representing the two levels in the Nelson and Narens (1990, 1994) framework. If an analogous task were conducted on a person, the response at the object-level would be to answer a question such as, "Who painted the Sistine Chapel?" Then, after giving the response, he or she would make the meta-level judgment about the correctness of the answer on a confidence rating scale. Several studies have shown that humans can make accurate confidence judgments (Koriat, Lichtenstein, & Fischhoff, 1980; Perfect & Hollins, 1996; Shaughnessy, 1979). A schematic of our task in the framework of Nelson and Narens is provided in figure 12.1B.

Our experiment was designed to test, in rhesus macaque monkeys, the existence of true meta-judgments (explicit or implicit) about their

previous cognitive responses. The monkeys were first asked to touch the longest of nine lines (analogous to a trivia question for a human). After making their response, they were asked for their judgment of certainty of their response by making a high or low bet. Unlike making an escape response, where there is only one response, the difficulty of our task was that the monkey had to remember the object-level response he had just made, but he would also have to connect that response to a subsequent meta-judgment. Furthermore, unlike the escape procedures, the benefit of asking for both object-level and meta-level responses is that we could tell on a given trial both whether the monkey knew the answer and how sure he was. Of course, a monkey confidence judgment could not be a simple rating of 0 to 100 as it would be for humans. Instead, the betting paradigm was used, in which the monkey could wager a lot on a response (in terms of reward), or play it safe and risk only a little.

EXPERIMENT 1: THE BETTING PARADIGM

Two male rhesus macaques (*Macaca mulatta*), named Ebbinghaus and Lashley, were tested in a chamber containing a touch-sensitive computer screen. On each trial, they were presented with nine lines positioned vertically on the screen, in a 3 × 3 array. Their task was a psychophysical task, in which they had to press the longest of the nine lines. Trials were either easy or difficult: On easy trials, one of the nine lines was noticeably longer than the other eight, which were all the same length. On difficult trials, all nine lines were the same length, but one line was arbitrarily designated as correct, so the monkeys were correct on one out of nine trials by chance. For each touch, a green border appeared around the item briefly. Then the screen cleared and the monkeys were asked to make a bet by touching one of two icons: a high-risk icon and a low-risk icon. Figure 12.2 displays the four possible outcomes that could occur for the two risk responses. If high risk was selected, the reinforcement was the gain of two tokens given a correct response on the line task, or the loss of two tokens given an incorrect response (the tokens were later exchanged for pellets). If the low-risk icon was touched, the reinforcement was always a gain of one token regardless of accuracy on the line task. The token reservoir was located on the bottom right-hand side of the screen. Tokens appeared to fly out of or into the reservoir with noticeable noises. Once the number of tokens in the reservoir reached or exceeded 12, the monkeys received two real banana-flavored food pellets and the token reservoir reset to 9 tokens. Using the tokens allowed us to calibrate the amounts of reward for each level of risk, so as to minimize bias toward one of the risk options. Toward the same end, when a monkey developed a bias toward one of the risk choices, that

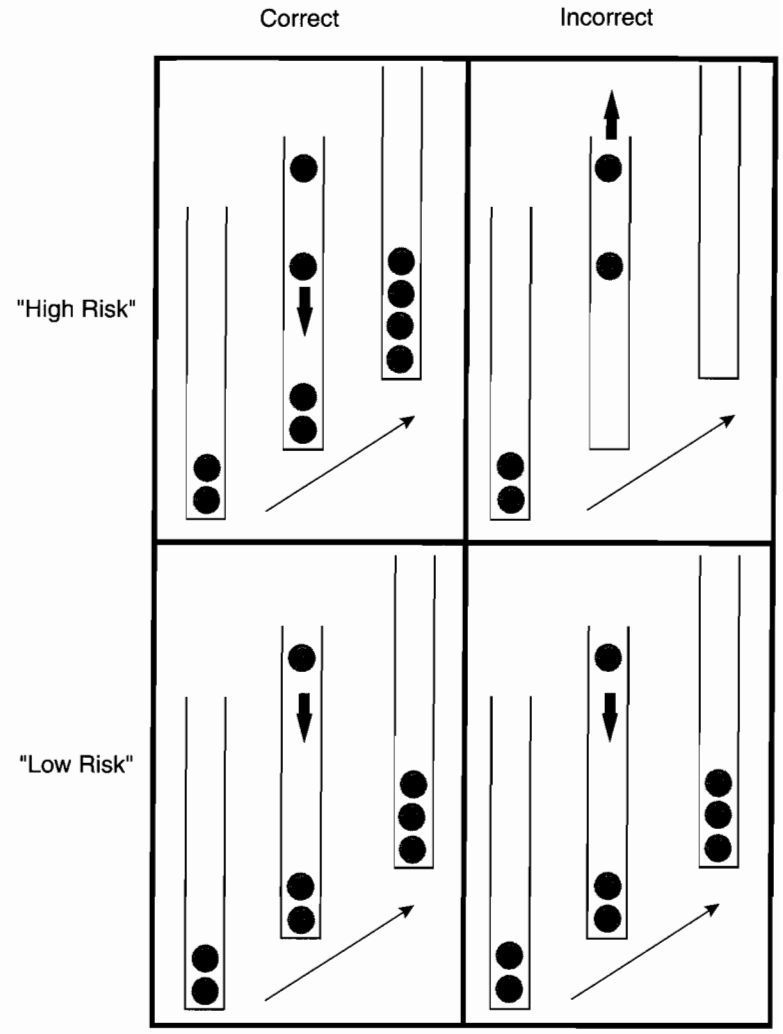

Figure 12.2 Contingency table showing the four possible outcomes of correct or incorrect response to the line task and high- and low-risk bets, starting with two tokens already present in the reservoir. When high risk was pressed, two tokens fell into the reservoir after correct responses, but flew out of the reservoir after incorrect responses. When low risk was pressed, one token fell into the reservoir regardless of whether the line response was correct or incorrect. Note that in Experiment 2, choosing high risk resulted in a gain or loss of three, rather than two, tokens. See color insert.

risk icon began to appear only after a delay, which ensured that the monkeys chose both levels of risk with some frequency. A sample trial is presented in figure 12.3A.

The hypothesis was that on easy line-discrimination trials, the monkeys would get the answer correct, presumably feel certain about it, and consequently bet high risk. On difficult (impossible) trials, though, they would (usually) get the answer incorrect, presumably feel uncertain, and bet low risk. Again, an advantage of this paradigm was that two responses were required: the line discrimination at the object-level, and a judgment of that response—the bet—at the meta-level.

Training

Typically, it is advantageous to reward animals immediately following a response. This keeps the animal motivated. Furthermore, immediate feedback is most easily associated to the one and only most recent response that has been made. Thus, an animal should have little trouble understanding what response is being rewarded. However, this was not the case in the current paradigm. Rather than immediate feedback, the monkeys learned that feedback occurred only after making two separate responses. Never before had these monkeys been asked to perform two separate tasks in a single trial. Essentially, after the monkey made his first response, the screen went blank and a new question appeared. The monkey's task was not only to make a second response but also to understand that the feedback he received not only depended on that second response but reflected two past responses—the bet response in addition to the earlier line response. We believe this was perhaps the most challenging aspect of the experiment. In addition, we required that the monkey remember four different contingency combinations and respond accordingly. Finally, the monkey needed to learn that accumulating a certain number of tokens on the computer screen resulted in a food reward after a criterion had been reached. Thus, because of the novelty of the paradigm for these participants, a substantial amount of training was involved prior to the actual initiation of the experiment. This included training of three distinct aspects: the main task, the tokens (including reward and punishment), and the betting icons.

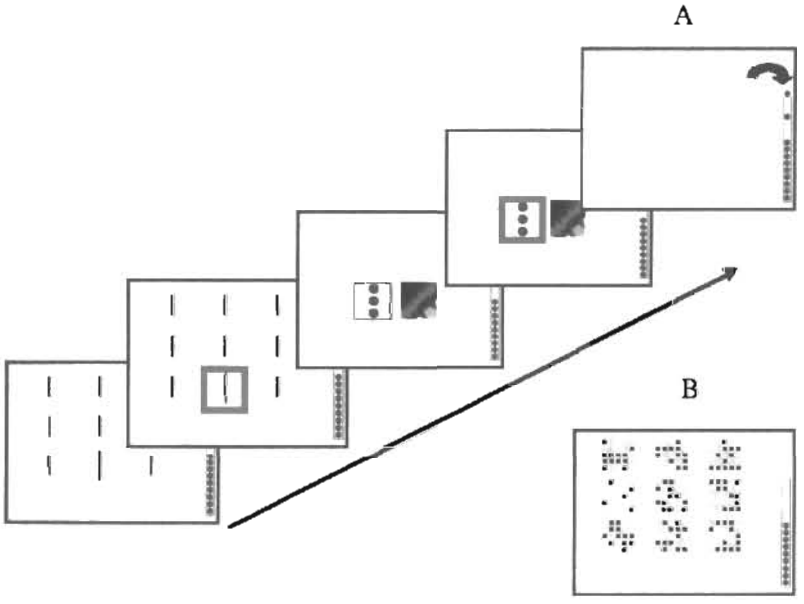

Figure 12.3 A. A sample trial for Experiment 1. In the first phase, nine lines appeared on the screen. A token reservoir appeared in the bottom right-hand corner of the screen, set at nine at the beginning of each session. The monkey's task was to press the longest line. For each touch, a green border appeared around the pressed item. Then the lines disappeared, and two risk icons appeared. The monkey's task was to report his confidence by betting either low or high risk. Once he made his bet, the token reservoir changed accordingly. In this sample trial, the monkey has bet high risk, and was correct on the line task. Thus, two tokens were added to the reservoir. B. Sample numerical stimuli for Experiment 2. The task was to press the stimulus with the most (Lashley) or the least (Ebbinghaus) items in them. Otherwise, the general methods remain the same as those in Experiment 1. See color insert.

The monkeys were first trained on the psychophysical task alone. At the start of training, Ebbinghaus and Lashley were presented with two vertical lines of different lengths on the screen. Their task was to press the longer of the two lines. A food pellet was earned only after correct responses. Both participants learned the task very quickly, after only a few sessions. Gradually, the number of lines presented for each question was increased, to three, then four, until they were presented with a total of nine vertical lines. All sessions were 20 minutes in length.

About a year after the beginning of training on the line task, a token reservoir was added on the bottom right-hand side of the screen during the entirety of each session. (During that year, we tried a similar betting procedure using food pellets as rewards and time-outs as punishments. Specifically, a high-risk response resulted in either a big food reward or a long time-out, whereas a low-risk response resulted in a small food reward or a short time-out. However, using this paradigm, both monkeys had a significant bias for pressing high risk. We thought that this might have been because the aversion to the time-outs was not nearly as salient as the attraction of the food rewards. We attempted to fix the problem by tweaking the contingencies—increasing the length of the high-risk time-out period—but there was no evidence that the monkeys understood the connection between the line task and the risk contingencies, and so we gave up.) Following each correct response to the line task, a token flew into the reservoir.

After each incorrect response, the token reservoir remained empty. At the start of token training, the criterion for receiving real food pellets was set at one token, so that whenever the monkey responded correctly, a token would fly into the reservoir and then would disappear at the bottom, simultaneously with the delivery of a food pellet. Gradually, the criterion was increased. Thus, the monkeys learned that an accumulation of these tokens was a positive feature that would eventually earn them food reward. At this time, the line trials were kept at a fairly easy level, so that the monkeys would not get frustrated and would continue to be motivated during each session. (We would experience motivation difficulties several times over the course of the research.) Then the number of starting tokens in the reservoir was varied. For example, for some sessions, the reservoir started at a level of three, with six as the criterion—meaning that in order to receive food, the monkeys needed to get three trials correct. We had the start number begin at a number greater than zero because the next step of training would be to introduce punishment, or the loss of tokens. We were also interested in having the monkeys connect the tokens to actual food pellets, which they did. In fact, throughout training, both Ebbinghaus and Lashley spent some time licking the tokens on the screen, and attempting to push the tokens down, as though they could force the tokens to come out as real food pellets. In the meantime, the trials were made gradually more difficult (the ratio of difficult to easy trials increased).

The next major training step was to introduce the punishment—in our paradigm, a loss of tokens—to the monkeys. We were not sure how aversive such a punishment would be. (We only hoped that this time, the punishment would be more effective than the time-out punishment that we had attempted to use prior to incorporating the token economy.) Never had any food pellets, tokens, or other rewards been "snatched away" from the monkeys prior to this experiment. Here, when an incorrect response was made, a token (only if there were already at least one in the reservoir) would fly up and out of the reservoir, making a "negative" noise, and disappear. Of course, by this time, the monkeys knew that fewer tokens in their reservoir meant they were farther away from reaching criterion or winning a food pellet. Thus, on the first session of punishment training, both monkeys seemed almost stunned, staring at the reservoir on the screen for a while and showing obvious frustration behaviors. Lashley, after a couple minutes of frustration, soon settled in again, performing some more trials and continuing on in training. Ebbinghaus, on the other hand, stopped working, forcing us to put him on a period of remedial training with easier trials without punishment for a while. Slowly, Ebbinghaus got back on track and learned to work

consistently even with the punishment. Although we were happy to know that the punishment using the token paradigm was adequately aversive for the monkeys, one future interest of ours is to investigate the differences in salience for rewards and punishments by manipulating the contingencies.

Following punishment training, both monkeys were trained to respond at a stable pace during 20-minute sessions, with a mixture of both easy and difficult trials. For each session, the reservoir began with 9 tokens, while the criterion was set at 12. Finally, the risk icons were introduced. Both icons were presented on the screen after the line response was made. One worry was that up until this point using the token procedure, the monkeys had always received feedback immediately after making their line response. This would be the first time they would not receive immediate feedback. The risk icons that we chose were two pictures, each symbolizing the number of tokens that would be risked if pressed. For example, the high-risk icon was a picture of a square with two tokens and the low-risk icon was a picture of a square with one token. The contingencies, set up like a logical gamble, were as follows: For betting high risk, they earned two tokens after a correct line response and lost two tokens after an incorrect line response. For betting low risk, they earned one token after a correct line response and lost one token after an incorrect line response. The monkeys soon learned that each of the pictures resulted in different contingencies. Whether they knew that it had anything to do with their previous line response was not yet known. In fact, again, both monkeys acquired an early bias for pressing high risk. This led us to increase the number of difficult trials during the session. We also changed the low-risk contingencies to always gaining a token regardless of previous line accuracy (changed from the previous low-risk contingencies of gaining a token when correct and losing a token when incorrect). When we still had difficulties in encouraging the monkeys to select the low-risk option more often, we changed the low-risk icon to an arbitrary picture (see figure 12.3A, the middle picture of the row of five) because we were concerned that the original low-risk icon had become unattractive. At the end of training, then, which was approximately 7 months after the tokens had first been introduced, the betting procedure was one in which a high bet meant risking two tokens, while a low bet was gaining one token regardless of previous line accuracy. Furthermore, due to the high-risk bias, a time delay was added prior to the appearance of one—the more biased one—of the risk icons (i.e., the more biased they were toward high risk, the longer they had to wait before the high-risk icon appeared). Starting from this point, we were able to obtain 41 subsequent sessions of data for both Ebbingaus and Lashley.

Results

The data were analyzed over a total of 41 sessions for each monkey. The average number of trials completed in each session was 87 and 60 for Ebbinghaus and Lashley, respectively (each session lasted 20 minutes). The mean percentage of easy trials answered correctly was 94% for Ebbinghaus and 88% for Lashley, and the mean percentage of difficult trials answered correctly was 12% for both monkeys (11.1% is chance with nine items to choose from). This is slightly higher than chance, but the difference was not significant for either monkey.

For each session, a phi correlation was computed between accuracy (correct vs. incorrect) and bet (high risk vs. low risk). We expected the monkeys to bet high risk on correct trials and low risk on incorrect trials, which would result in positive correlations. The correlations, blocked in sessions of four, are presented in figure 12.4. As shown, the correlations were generally positive.

In general, although Ebbinghaus took a little longer to achieve high correlations, both monkeys were able to report stable positive correlations by the end of training. For all 41 sessions, the mean correlations

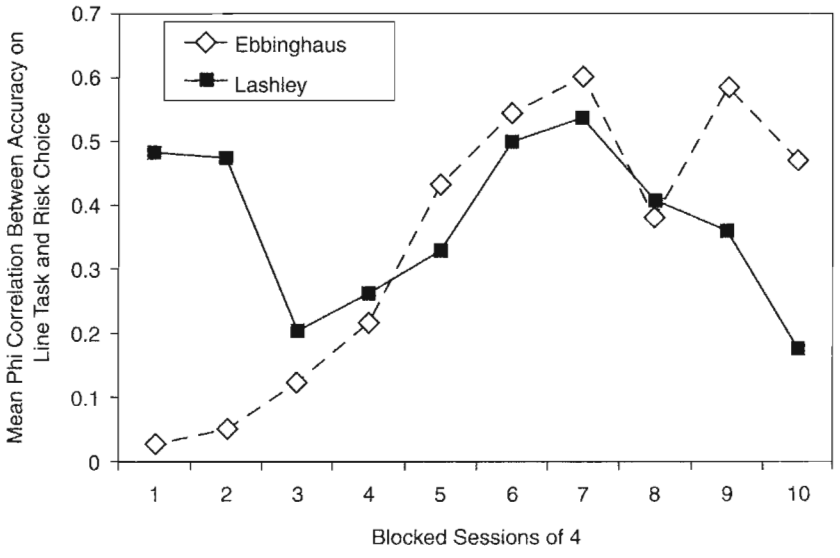

Figure 12.4 The mean phi correlations between accuracy of the object-level line task and the meta-level risk choice for the final 41 sessions of training, in blocks of four. A positive correlation indicates that on trials that were correct, the monkeys more often bet high risk, whereas on trials that were incorrect, they more often bet low risk.

were significantly above zero for both Ebbinghaus, $M = 0.33$, $t(40) = 8.43$, $p < .0001$, and Lashley, $M = 0.37$, $t(40) = 13.37$, $p < .0001$. We also calculated the percentage of correct and incorrect trials on which the monkeys chose high risk. As figure 12.5A indicates, both monkeys chose high risk significantly more often after correct trials than after incorrect trials—for Ebbinghaus, $t(40) = 8.44$, $p < .0001$; for Lashley, $t(40) = 13.31$, $p < .0001$.

We also calculated reaction times (RT) for the difficult and easy trials (for all RT analyses, we excluded trials over 10 seconds long). The results showed that both monkeys took longer to respond to the line-discrimination task on difficult trials than on easy trials. Ebbinghaus took an average of 1.24 seconds to respond on difficult trials, but only 1.07 seconds on easy trials, which significantly differed from each other, $t(40) = 3.19$, $p < .01$. Lashley also took significantly longer to respond on difficult trials, averaging 2.00 seconds as compared to 1.66 seconds on easy trials, $t(40) = 2.55$, $p < .05$. In general, a longer response time could be interpreted as resulting from more uncertainty.

In summary, using a betting paradigm that included both an object-level task and a meta-level task, it appeared that both Ebbinghaus and Lashley were able make accurate confidence judgments. When they responded correctly on the object-level task, they were able to report that they were certain by tending to bet high risk. Similarly, after responding incorrectly, they bet low risk more often. If they were human, this chapter could now conclude without much controversy by saying that "Ebbinghaus and Lashley were able to express feelings of certainty and uncertainty about their cognitions, thus showing metacognitive abilities."

Discussion

However, the reader may be curious, as we were, to know whether these responses were based on judgments resulting from direct or explicit access of internal representations, or from inferential mechanisms. Regardless of what the basis was, both of these mechanisms, according to the human literature, would indicate that the monkeys possess metacognitive abilities. Given this, several questions arise. To begin with, perhaps the monkeys relied on cues that did not require any feelings of uncertainty. For example, the monkeys could have been relying on their own response times to guide their bets. This seems plausible, since easy and difficult trials resulted in a difference in response latencies—they might have simply bet low risk when they observed that they had taken a long time to make their response during the line task, and high risk when they had been fast. Alternatively, they could have relied on pattern recognition to make their bets. For example, they might have responded low risk when they recognized

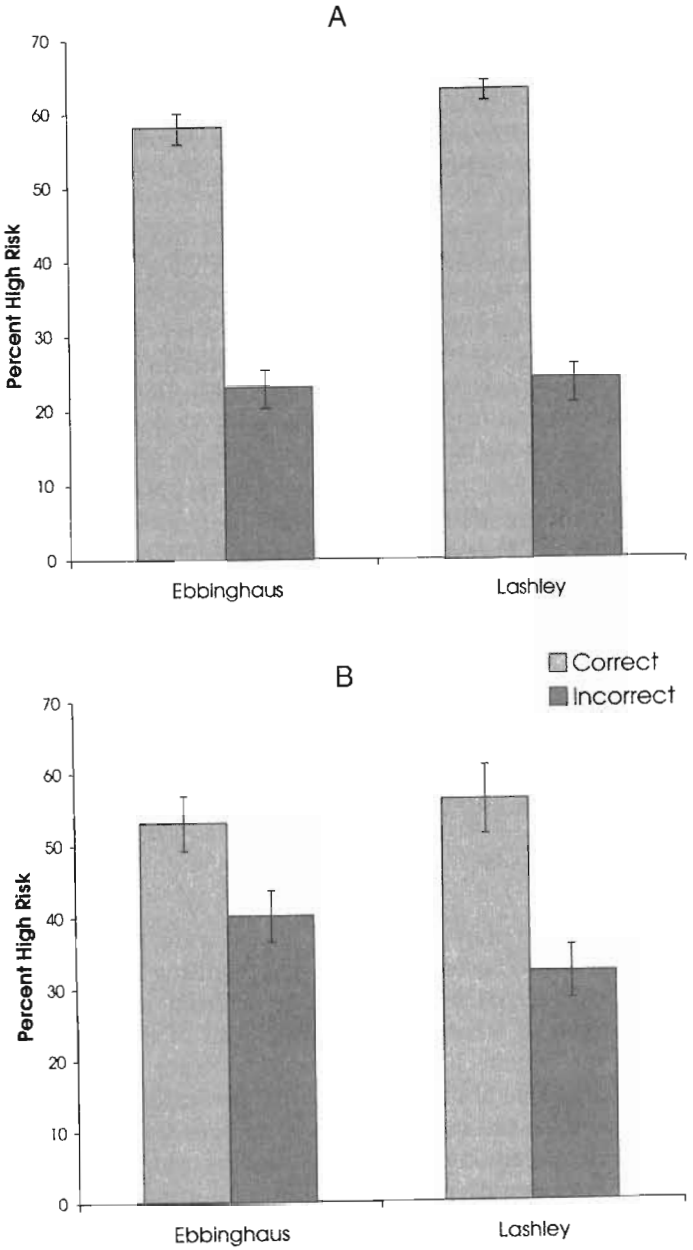

Figure 12.5 Mean percentage of correct and incorrect trials on which high risk was chosen in (A) Experiment 1 and (B) Experiment 2. These means were calculated for each monkey over the final 41 sessions for the line task in Experiment 1 and over the first 4 sessions for the numerosity task in Experiment 2.

that all of the lines were the same length, without actually feeling uncertain. If either of these was the case, then one hesitates to say that the monkeys were responding based on explicit mechanisms.

It is interesting to note that these two possibilities are quite similar to two inferential mechanisms in relation to human metacognition: retrieval fluency and cue familiarity. As described at the beginning of this chapter, in retrieval fluency, the faster one arrives at an answer, the higher the judgment will be. The cue familiarity hypothesis states that meta-judgments are based on how recognizable the cue is (Metcalfe, 1993a). For example, a person may give a high judgment for the question "What was the name of Batman's butler?" because of high familiarity in the subject of Batman and not because of any actual memory of the answer. Cue familiarity as a metacognitive cue has been supported by many experimental results (Glenberg et al., 1987; Metcalfe, 1993a, 1993b; Metcalfe et al., 1993; Miner & Reder, 1994; Reder, 1987; Reder & Ritter, 1992; Schwartz & Metcalfe, 1992). In both retrieval fluency and cue familiarity, metacognition is inferred from factors other than the explicit assessment of memory itself. Still, in the above-mentioned literature, when these judgments were found to be based on such inferential mechanisms, they were classified as metacognitive judgments. Based on this, if Ebbinghaus and Lashley were using cue-based pattern recognition or their reaction time as a basis for their uncertainty judgments, one reaction is to say that they still do possess metacognitive abilities. Furthermore, they would be a lot like humans if they did either of these.

To investigate these possibilities, we first explored the question of whether retrieval fluency predicted risk choice for our monkeys. To do so, we calculated mean point-biserial correlations between response time to the line-discrimination task and the bet choice. If the monkey bet low risk when he was slow on the line task and bet high risk when he was fast, the mean correlation across the 41 sessions would be negative, and it was for both Ebbinghaus, $M = .12$, $t(40) = 4.43$, $p < .0001$, and Lashley, $M = .17$, $t(40) = 5.64$, $p < .0001$. Based on this finding, it seems that response latencies could have helped guide the monkeys confidence judgments. This opens up the possibility that the monkeys behavior can be explained without positing that they made judgments based on explicitly assessing internal representations. However, we had not yet investigated whether there was any evidence that the monkeys might have, in fact, been making their risk judgments based on something internal.

To explore the question of whether the monkeys were monitoring an internal representation, we reanalyzed the data, factoring out the effect of both trial difficulty and RT. First, we analyzed only the easy trials, which were all equally difficult, meaning the monkeys could not use difficulty as a cue, which even for a human makes meta-judg-

ments much more difficult. We then computed Pearson correlations (although the Pearson correlations are standard for continuous variables, we used them here because there was no standard way of computing partial correlations between binary and continuous data) between accuracy of the line task and risk choice, partialing out the effect of RT, separately for each session. A positive correlation would indicate that the judgments were attributable to mechanisms other than those based on trial difficulty or RT. The mean correlations were significantly positive for both Ebbinghaus, $M = 0.07$, $t(35) = 2.67$, $p < .05$, and Lashley, $M = 0.20$, $t(29) = 4.24$, $p <; .001$, although more so for Lashley than for Ebbinghaus. These results reveal that both monkeys might be making meta-judgments based on something more than implicit cues such as assessments of internal representations.

Another way to investigate whether the monkeys were not merely basing their confidence judgments on external pattern recognition, relying on features specific only to the line task, was to present the monkeys with a novel question, a transfer task. Transfer tasks are widely used, typically to ensure that the tested skill—in this case, the reporting of accurate confidence judgments through bets—has not been learned as an automated response to the external stimulus that was presented, and can be generalized to other tasks. Thus, in Experiment 2, we presented an experiment similar to Experiment 1, except that the line task was replaced with a new task. First the monkeys were trained on the new task until performance was stable, and then the risk choices were added. If they could make appropriate risk choices immediately, we reasoned that their ability to make confidence judgments had generalized and was not based on features specific to the line task in Experiment 1.

EXPERIMENT 2: TRANSFER

The purpose of Experiment 2 was to test whether the ability to make accurate meta-level confidence judgments would generalize to a numerosity task in which the monkeys had to differentially choose one correct picture out of nine presented on the screen (see figure 12.3B for a sample stimulus). Both Ebbinghaus and Lashley had previous experience with a numerical task in which pictures containing varying numbers of items (e.g., a yellow square containing 12 black dots) were presented simultaneously. Ebbinghaus had learned to press the pictures in an ascending order, and Lashley had learned to press the pictures in a descending order. We modified our task so that the monkeys were presented with nine pictures, each containing some number of items. The monkeys had to press only the one picture that contained either the most items (in Lashley's case) or the fewest items (in Ebbinghaus's case). For both monkeys, each of the eight incorrect

choices contained the same number of dots. As in Experiment 1, trials varied in difficulty, but instead of only two levels of difficulty, there were four levels (none of them being impossible as one was in Experiment 1): On the very easy trials, one of the nine pictures consisted of noticeably more (or fewer) items than the other eight choices. As the trials became increasingly difficult, the correct answer became less and less distinguishable from the distractors. All sessions were 20 minutes in length. Once the monkeys had been trained on this paradigm (which took three sessions for Ebbinghaus and seven sessions for Lashley), the risk choice was added. After the numerical response was made, the computer screen cleared and the monkeys were asked to make their bet. The reinforcement contingencies were slightly modified from Experiment 1, in which the monkeys gained or lost two tokens when they bet high risk. Here, in order to increase the number of pellets the monkeys received, the contingency was changed to a gain or loss of three tokens following high-risk choices. Our hope was that, starting with the very first sessions, both Ebbinghaus and Lashley would behave metacognitively, by risking more when they were correct than when they were incorrect. These results would suggest that the monkeys had understood the meaning of the bet choices, rather than choosing them in response to patterns of the line task.

Results and Discussion

In Experiment 1, we were interested in investigating whether the monkeys acquired a metacognitive skill through training on the betting paradigm. Here, though, we were not interested in whether the monkeys could acquire the metacognitive skill. Rather, our interest lay in the very beginning of training, to examine whether immediate transfer of the betting skill occurred for a new task. Thus, in Experiment 2, we analyzed the first four sessions of training. To analyze the relationship between accuracy and risk, we again calculated phi coefficients, which allowed us to compute Fisher's exact p-values for a small number of observations. Both monkeys showed positive correlations almost immediately. For Lashley's first four sessions combined, $\varphi = 0.24$, $p < .0001$. For Ebbinghaus's first four sessions combined, $\varphi = 0.14$, $p < .05$. We also calculated the percentage of trials on which high risk was chosen separately for correct and incorrect trials, again combining the first four sessions, as shown in figure 12.5B. As in Experiment 1, high risk was chosen more often on correct trials than on incorrect trials.

The transfer task showed that both monkeys began making good risk decisions directly upon being introduced to their new task. They did not need training to make confidence judgments on the new task—unlike in the line/risk task, which took many months to learn. Instead, they showed enough flexibility to generalize the risk response

to the new task. This makes sense if they were making their risk responses as meta-judgments. Such judgments about cognitions should occur for virtually any cognition, and it appears that what the monkeys had learned was how to respond with accurate *meta*-cognitions.

We were aware of the fact that although the transfer results make a stronger case for us, there are still similarities between the numerical task and the line task, the most obvious being that both involve some trials where all the stimuli look very similar, and other trials where one stimulus is quite different than the others. Thus, the monkeys could have succeeded by responding to the risk choices based solely on whether the stimuli were the same or different, in both experiments. We answered this criticism earlier by showing that risk choices were appropriate even when only easy trials (which are all the same) were included in the analysis. But it might be argued that on incorrectly answered easy trials, the monkeys simply did not notice the correct answer, and therefore thought that all of the stimuli were the same. In that case, even on easy trials, they would choose low risk and get the answer incorrect without necessarily relying on uncertainty. In short, one of the concerns of the present tasks is that they are both perceptual tasks that may not have required internal representations. So, although we think our findings are convincing, the best way to answer criticisms about stimulus cues is to use a task where nothing about the stimuli themselves makes it possible to make the risk choice successfully. One way to address this concern would be by using the betting paradigm with a memory task (e.g., Hampton, 2001; Smith et al., 1998; chapter 11, this volume), not a perceptual task.

Tasks that rely on the monkeys' memory make it difficult for a meta-judgment to be made on the basis of external cues. For example, Smith et al. (1998) used a serial probe recognition task, in which monkeys were shown four pictures in sequence and then a probe, and had to indicate whether the probe had been in the preceding list. If the probe had been presented in the list, the participants were to press the probe. If the probe was new, they were to press a second option that represented the idea of "new." If they were unsure whether the probe was old or new, they were to press a third option, the escape response. In this paradigm, when the monkey is faced with the probe, the only thing that indicates to him whether the trial is easy or difficult is the strength of his memory trace. The appearance of the stimuli offers no hints. Combining the serial probe recognition task with the current confidence-judgment task would have the advantage that on each trial the monkeys would make a response before placing their bet. As a result, we would know which trials were difficult for them. Also, when using only the escape response, if the item is very familiar or very unfamiliar, the monkeys will have a strong urge to respond either *old* or *new*. On difficult trials, if they have no strong urge they might simply respond randomly, or choose the escape response by default, and as a result choose the escape response relatively frequently. Confidence judgments do not allow such a strategy. Neither does the match-to-sample task used by Hampton (2001; chapter 11, this volume): By forcing the monkey to choose between high and low risk and then perform the task, he ensured that his monkeys could not choose low risk randomly or by default.

Memory tasks avoid certain criticisms, but there is a deeper reason to favor them. Our monkeys' estimations of size and number are cognitions, and thus their performance fits the traditional definition of metacognition. However, metacognitive research has been traditionally associated with judgments and assessments about memory. Thus, our next step is to test our monkeys' ability to make confidence judgments using a memory task, namely, one similar to the serial probe recognition task.

CONCLUSION

The current definition of metacognition is associated with many different functions, some of which (e.g., theory of mind) go far beyond making smart risk choices. For example, Nelson (1992) defines it as a form of self-reflective consciousness, in which we are able to consciously observe the workings of our own minds. This type of metacognition is thought to be a high-level process that entails privileged access into one's own internal mind and may be impossible to observe in nonverbal animals. However, another definition of metacognition takes a different view—that many of the mechanisms involved in making metacognitive judgments need not be open to conscious awareness (Cary & Reder, 2002; Reder & Schunn, 1996). According to this view, we make metacognitive judgments constantly and without explicit knowledge of them. These implicit metacognitions can be based on internal factors, but may also be based on ongoing external factors that we are not aware of.

A link can be drawn between direct-access mechanisms and conscious metacognition, in that both involve inward reflection. Feelings of uncertainty that are inferential and based on cue familiarity (for example) do not necessarily require self-reflection any more than a simple familiarity judgment does. Yet it is also clearly possible to feel uncertain based on the familiarity of a cue and, at the same time, be quite aware of that feeling. Likewise, it is also the case that an animal (or human) might have a feeling of uncertainty based on a direct access mechanism and not be aware of the process. In fact, some researchers have claimed that people have little or no direct introspective access to mental processes such as those affecting judgments (Koriat, 1993; Nisbett & Bellows, 1977; Nisbett & Wilson, 1977). For

example, Koriat (1993) provides evidence supporting an accessibility mechanism of metacognition which states that, although people may access internal representations and base their judgments on those representations, they have no direct awareness of the accuracy of the internal representations. Instead, people's meta-judgments are based on both correct and incorrect pieces of information that were purportedly retrieved from an internal memory trace.

Given these issues, although it may seem that there could be a correspondence between direct access and consciously aware metacognition, it cannot be confirmed. If the process of accessing internal information from memory is not necessarily open to conscious awareness, then we see no reason why the result of the process has to be either. Thus, neither explicit mechanisms nor implicit mechanisms are "more metacognitive" than the other. Regardless of how humans, or our monkeys, made their judgments, it does not take away from the fact that the meta-judgments were made.

In this chapter, we have presented evidence that two rhesus macaque monkeys were able to display metacognitive abilities by reporting high- and low-risk bets about their own cognitions. If metacognition is taken to mean self-reflective consciousness, then our conclusion is, of course, endlessly arguable. However, we believe that the monkeys were making the type of metacognitive judgments that occur in people all the time, with or without awareness. We readily concede that because they cannot verbally express their judgments, we do not have evidence of awareness of their judgments. If we did, we would then have evidence of *meta*-metacognition (which people express when they verbalize their metacognitions). The results from the present experiments allow us to conclude that our monkeys are able to reflect upon their cognitions, indicating that they possess at least one level of metacognition. And that suffices for granting them metacognitive abilities.

REFERENCES

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81,* 126–131.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68.

Cary, M., & Reder, L. M. (2002). Metacognition in strategy selection: Giving consciousness too much credit. In M. Izaute, P. Chambres, & P. J. Marescaux (Eds.), *Metacognition: Process, function, and use* (pp. 63–78). New York: Kluwer.

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development, 24,* 15–23.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119–136.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences U.S.A., 98,* 5359–5362.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208–216.

James, W. (1890). *Principles of psychology.* New York: Holt.

Janowsky, J. S., Shimamura, A. P., & Squire, L. R. (1989). Memory and metamemory: Comparisons between frontal lobe lesions and amnesic patients. *Psychobiology, 17,* 3–11.

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

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118.

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

Metcalfe, J. (1993b). Monitoring and gain control in an episodic memory model: Relation to the P300 event-related potential. In A. F. Collins & S. E. Gathercole (Eds.), *Theories of memory* (pp. 327–353). Hove, UK: Lawrence Erlbaum.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 851–864.

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing.* Cambridge, MA: MIT Press.

Miner, A. C., & Reder, L. M. (1994). A new look at feeling of knowing: Its metacognitive role in regulating question answering. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 47–70). Cambridge, MA: MIT Press.

Nelson, T. O. (1992). *Metacognition: Core readings.* Boston: Allyn and Bacon.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General, 113,* 282–300.

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. 125–141). 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.

Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of Personality and Social Psychology, 35,* 613–624.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84,* 231–279.

Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 10,* 371–382.

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

Reder, L. M. (1996). Different research programs on metacognition: Are the boundaries imaginary? *Learning and Individual Differences, 8,* 383–390.

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.

Reder, L. M., & Schunn, C. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 45–77). Mahwah, NJ: Lawrence Erlbaum.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin and Review, 1,* 357–375.

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1074–1083.

Shaughnessy, J. J. (1979). Confidence-judgment accuracy as a predictor of test performance. *Journal of Research in Personality, 13,* 505–514.

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.

Shimamura, A. P., & Squire, L. R. (1986). Memory and metamemory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 452–460.

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., & Erb, L. (1995). The uncertain response in the bottlenosed dolphin (*Tursiops truncates*). *Journal of Experimental Psychology: General, 124,* 391–408.

Smith, J. D., Shields, W. E., Allendoerfer, K. R., & Washburn, D. A. (1998). Memory monitoring by animals and humans. *Journal of Experimental Psychology: General, 127,* 227–250.

Smith, J. D., Shields, W. E., Schull, J., & Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition, 62,* 75–97.

Tulving, E. (1994). Foreword. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. vii–x). Cambridge, MA: MIT Press.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. In P. H. Mussen & M. R. Rosenzweig (Eds.), *Annual review of psychology* (pp. 437–484). Palo Alto, CA: Annual Reviews.

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free-recall learning. *Journal of Experimental Psychology, 71,* 673–679.