
Biology and Philosophy **19:** 633–653, 2004.

Do dolphins know their own minds?

DEREK BROWNE
Philosophy Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand; E-mail: derek.browne@canterbury.ac.nz

Key words: Dolphin, Self-knowledge, States of mind

Abstract. Knowledge of one's own states of mind is one of the varieties of self-knowledge. Do any nonhuman animals have the capacity for this variety of self-knowledge? The question is open to empirical inquiry, which is most often conducted with primate subjects. Research with a bottlenose dolphin gives some evidence for the capacity in a nonprimate taxon. I describe the research and evaluate the metacognitive interpretation of the dolphin's behaviour. The research exhibits some of the difficulties attached to the task of eliciting behaviour that both attracts a higher-order interpretation while also resisting deflationary, lower-order interpretations. Lloyd Morgan's Canon, which prohibits inflationary interpretations of animal behaviour, has influenced many animal psychologists. There is one defensible version of the Canon, the version that warns specifically against unnecessary intentional ascent. The Canon on this interpretation seems at first to tell against a metacognitive interpretation of the data collected in the dolphin study. However, the model of metacognition that is in play in the dolphin studies is a functional model, one that does not implicate intentional ascent. I explore some interpretations of the dolphin's behaviour as metacognitive, in this sense. While this species of metacognitive interpretation breaks the connection with the more familiar theory of mind research using animal subjects, the interpretation also points in an interesting way towards issues concerning consciousness in dolphins.

Introduction

There are several kinds of self-knowledge. One kind is psychological self-knowledge, knowledge of one's own states of mind. We tend to assume that psychological self-knowledge is one of those higher mental achievements that is beyond the reach of most or even all nonhuman animals (henceforth just 'animals'). Some recent research in animal psychology challenges this assumption. The research provides evidence that some animals have some knowledge of their own states of mind. One very interesting feature of this research is that the species in question is not one of the more usual subjects for research into higher cognition, namely, the great apes. The species is the bottlenose dolphins (*Tursiops truncatus*). Dolphins are phylogenetically distant from hominids and great apes. Whereas the closest common ancestor of humans and chimpanzees lived around 6–7 million years ago, the closest common ancestor of humans and cetaceans lived somewhat more than 65 million years ago. Dolphins are also large-brained animals. When measured for relative brain size, a few species of dolphins, including the bottlenose

dolphin, turn out to be second only to humans in the size of their brains, and to score well above the great apes (Herman 2002).

Intentional ascent

In philosophy, the main conceptual tool for the analysis of psychological self-knowledge is the concept of intentional ascent. Intentional ascent leads a subject from lower to higher levels of intentionality (Dennett 1987). If a thinking subject is not thinking about intentional states but is thinking about something else – anything else – then that subject is in a first-order intentional state. If a subject is thinking about intentional states, then that subject is in a higher-order intentional state. Higher-order intentional states are other-directed when they are about the intentional states of another thinking subject. Higher-order intentional states are self-directed when they are about the subject's own, lower-order intentional states. This is the basic picture, but there are conceptual nuances.

There is higher-order intentionality if (and only if) there is a mental representation of a mental representation. A mental representation consists (in Brentano's canonical model) of a propositional content and a psychological attitude towards that content: a belief that p, a desire that q, a hope that r, and so on. There is metarepresentation of a belief that p only if both the content, p, and the attitude, believing, are represented in the higher-order state. If only the content and not the attitude is represented, then there is no metarepresentation: there is just a conjunction of two lower-order psychological states which have the same content. A psychological episode in which my belief that p causes me to be pleased that p is a first-order episode, notwithstanding that it involves two propositional attitudes with the same content. Higher-order intentionality occurs only when a propositional attitude takes another propositional attitude for its content. The simple rule for counting levels of intentionality is to count the psychological verbs (verbs of propositional attitude) in a third-person psychological attribution. For example: 'S thinks that p' refers to a first-order intentional state; 'S thinks that he (S) doubts that p' refers to a second-order intentional state; and so on. There is a potential trap hereabouts: when counting levels of intentionality in a Dennett-style soliloquy (Dennett 1987), the quotation marks that contain the soliloquy must be counted as introducing first-order intentionality (see below).

There is another nuance to notice. The canonical model of intentional ascent refers to states, all of which are intentional. But it sometimes happens that a thought is directed on a nonintentional psychological state: for example, S thinks that he (S) is experiencing a sensation of some specific kind. There is no question here of a higher-order representation of a lower-order propositional content, for there is no lower-order content (on the reasonable assumption that sensations do not have semantic content). But the thought is still a higher-order thought, at least in the sense that it represents another mental state. So

Dennett's famous lapwing soliloquy, 'all of a sudden I feel this tremendous urge to do that silly broken-wing dance' (Dennett 1987), refers to a higher-order thought about another mental state, an 'urge', which itself lacks intentional content. The second-order thought is introduced by the quotation marks: the lapwing is talking to itself. So this case is not an example of first-order mental activity, though it is an example of a propositional attitude which takes as its content another mental state, one which itself lacks semantic content. There can be higher-order thoughts in the absence of higher-order intentionality. Such nuances matter, when 'inflationary' and 'deflationary' interpretations for animal behaviour are in the balance.

The distinction between first-order thinkers and higher-order thinkers is important for several reasons. Here are two.

(1) An animal that cannot, in its own thinking, represent mental acts, is an animal that cannot be a mentalist, either about others or about itself. It cannot adopt the intentional stance, cannot interpret behaviour intentionally, cannot use intentional interpretations to predict the behaviour of others. The capacity for the intentional stance arguably played a critical role in the evolution of high intelligence among humans. Tomasello (1999) argues that the capacity to read the intentions of others is necessary for genuine learning by imitation to occur. In turn, imitation learning is the engine of cumulative cultural evolution. We got to where we are now through a co-evolutionary process, in which cognitive complexity and cultural complexity bootstrapped each other into existence.

(2) Intentional ascent might be the foundation of consciousness. There are several theories of consciousness currently on offer which associate the phenomenon with higher-order intentionality. The root idea is that a first-order mental state is a conscious mental state if and only if it is the intentional object of a suitable higher-order mental state. And if we can show that some variety of consciousness is cognitively constituted, then we might also be able to make progress on a problem that is believed by many to be utterly intractable: the question of animal consciousness. There will be evidence for consciousness in animals if there is evidence for the right sorts of higher-order cognition in animals.

Metacognition

Psychologists usually talk about metacognition rather than intentional ascent (Metcalfe and Shimamura 1994; Smith et al. 2003). This is not merely a semantic difference, because the common psychological model of metacognition is different in important respects from the philosophical model. The psychological model is functional. On this model, metacognitive states or processes are distinguished by the special functions they have in cognition. Specifically, a state is metacognitive if it has the function of monitoring other cognitive states or processes, or if it has the function of controlling other cognitive states or processes. Both functions, monitoring and control, are

understood in terms of information flow. In monitoring, there is an information flow from the 'object-level' to the 'meta-level', and in control, there is an information flow in the other direction. In virtue of the information flow from object-level to meta-level, models are generated at the meta-level of states and processes at the object-level. These models (or 'partial models') at the meta-level are available for use to regulate object-level processes. By way of illustration, a process of learning in a human subject might involve, at the start of the process, a judgement about the ease of the learning task. This judgement is the product of monitoring. On the control side, the judgement might be used in the allocation of cognitive resources. Later on, a feeling-of-knowing judgement, on the monitoring side, might lead on the control side to the termination of the learning process. (See Figure 5 in Nelson 1996, reprinted as Figure 1 in Smith et al. 2003).

This functional model of metacognition is broader than the philosopher's model based on intentional ascent. For one, it does not specifically require higher-order intentionality, since it does not specify that both lower-order psychological attitudes and their contents are represented at the higher level. As a result, it might be accused of blurring the important difference between higher-level processing and deeper-level processing that does not ascend above the first order of intentionality. If a representation in one cognitive processing system is transmitted to another cognitive system for further processing, then there is information flow between two cognitive systems, and the generation, in the second system, of a 'partial model' of the state of the first system (some features of the first-order state are represented in the second-order state). But further, first-order processing is not the same thing as higher-order processing, processing at a metacognitive level. A purely functional concept of metacognition needs to be enriched with a concept of metarepresentation in order to deliver higher-level intentionality.

There are also some advantages to the functional model. It copes easily with cases in which object-level states are not intentional, as in a thought about a sensation. It also copes with cases in which the object-level state is intentional, but in which the content of this state is abnormal in some respect. The feeling-of-knowing cases, to be discussed shortly, are an instance, as are the states of 'uncertainty' that are attributed to dolphins and other subjects in the uncertainty monitoring experiments, to be discussed more fully later.

The two models, one based on the concept of higher-order intentionality, the other on functional criteria, are not equivalent. But in contexts where the differences do not make a salient difference, I will refer simply to metacognition.

Experimental research on metacognition

My main topic is the research programme on higher-order cognition in dolphins. But first I will locate the dolphin research in the context of metacognitive research more generally. Some of this research is on humans, some on

nonhuman animals. Some research is on self-directed metacognition: a subject's knowledge of his or her own cognitive states. Some research is on other-directed metacognition: one subject's knowledge of another subject's cognitive states. The following are examples of research in each of these four categories:

(1) Other-directed metacognition in human subjects: developmental investigation into the stages through which human children acquire the capacity to attribute false beliefs to others (Carruthers and Smith 1996).
(2) Other-directed metacognition in animal subjects: investigation of the capacity for intentional deception in nonhuman primates (Whiten and Byrne 1988; Heyes 1998).
(3) Self-directed metacognition in human subjects: investigation of the reliability of the feeling of knowing (the tip-of-the-tongue phenomenon: Metcalfe and Shimamura 1994).
(4) Self-directed metacognition in animal subjects: investigation of uncertainty monitoring in dolphins.

(1) The false belief task is designed to discover the age at which young children acquire the capacity to attribute mental states to others. Children under about 4 years of age cannot grasp the possibility that another subject might have a belief that is false. False belief is a form of misrepresentation, representation of what is not the case. A young child who does not grasp the possibility of misrepresentation is not yet fully competent with the concept of representation. The false belief test looks for the ability to attribute mental states to others. It does not test for the ability to attribute mental states to oneself. The false belief task is also limited to subjects who are able to answer questions. So this procedure cannot be used without modification to test for theory of mind competence in nonhuman animals. But analogous procedures can be used. The procedures are analogous in that they depend on representational error. Can animals exploit false beliefs in other animals? Better still, do animals act so as first, to induce false beliefs in others, and then, so as to exploit those false beliefs?

(2) The research into intentional deception in primates investigates whether primates can intend to induce false beliefs in others. If so, then they are able to attribute beliefs to others. Various tests for intentional deception in primates have been attempted, but the results are inconclusive (Heyes 1998). There is some evidence for the existence of a capacity for intentional deception. The following story about Kanzi the bonobo chimp is particularly engaging, as well as illustrating some of the difficulties of this kind of research. Kanzi somehow managed to get hold of the key to his own enclosure, and promptly hid it. When Sue Savage-Rumbaugh started a search for the key, she invited Kanzi to help. Kanzi went through the motions of helping, pretending to search for the lost key. Later on, when things had quietened down, Kanzi retrieved the key and used it to escape from the enclosure (Whiten 2000). It looks as if Kanzi

intended to produce a false belief in Savage-Rumbaugh. Kanzi intended her to believe that the key was really lost: a second-order intention. Or even, more flamboyantly, Kanzi intended her to believe that he, Kanzi, did not know where the key was: a third-order intention. Is there a first-order explanation for Kanzi's behaviour? Yes there is: there usually is. Perhaps Kanzi just forgot that he had hidden the key and joined diligently and sincerely in the search for it. Only later did he remember hiding it.

Three different explanations have been found for Kanzi's behaviour, in two of which he is practising intentional deception. Which of the three explanations is most likely to be true? Animal psychologists often cite Lloyd Morgan's Canon at this stage of the inquiry. The Canon says that a higher-order psychological explanation must never be preferred if a lower-order explanation is also available (Morgan 1894; Sober 1998). Morgan's own statement of the principle presupposes a ranking of psychological states into higher and lower types. There is no need to rehearse the problems with qualitative distinctions like this. Yet in spite of its shaky rationale, Lloyd Morgan's Canon functions, for many in animal psychology, as an authoritative principle, such that its simple citation, unaccompanied by any reasons for accepting it as sound, gets to count as a substantive move in debates about the proper interpretation of animal behaviour.

There is, however, one context in which a distinction between higher and lower states has some application, namely, in relation to levels of intentionality. A higher-order intentional interpretation is more onerous than a lower-order intentional interpretation, in at least two ways. First, a lower-order interpretation attributes fewer cognitive states: it is simpler in that sense. Second, it is plausible that higher-order cognition is harder work than lower-order cognition, and so is less common, both in the mental life of an individual and in the phylogenetic distribution of the relevant capacities among the various taxa of animals. These considerations lend support to a methodological principle that advises us to prefer interpretations that invoke lower orders of intentionality to higher-order interpretations, all else being equal. This advice is far from decisive, however. Deflationary interpretations of complex behaviour reduce the number and the explanatory power of any 'hidden variables' that intervene between stimulus and behavioural response, at a cost sometimes of greater explanatory complexity than is found in alternative, higher-order interpretations of the same behaviour. Lloyd Morgan's Canon is not a principle of parsimony, because considerations of parsimony can favour higher-order explanations. However, in the context of metacognitive research on animal subjects, it is important to be open to the possibility that the behaviour can also be explained at a lower-order of intentionality, especially where such explanations do not strain credulity for other reasons. (I return to the question of Lloyd Morgan's Canon below).

I have described very briefly to two research programmes into metacognition: the false belief task provides a test for other-directed metacognitive knowledge in human infants, and the intentional deception research looks for

evidence of other-directed metacognitive knowledge in primates. I now describe, again very briefly, two research programmes which investigate self-directed metacognition, the first in humans, the second in dolphins.

(3) Sometimes I feel that I know the answer to a question but cannot quite get it: the answer is on the tip of my tongue. I am sure that I know the name of the Italian city in which Romeo and Juliet is set, even though I am not able for the moment to retrieve the name. The fact that an item of information is present in my long-term memory is a first-order cognitive fact about me. At the metacognitive level, I possess partial information about this first-order fact: I know that the information is present in memory even though I do not know its full content. In the metacognitive literature, these cases are known by the unlovely acronym FOK (Metcalfe and Shimamura 1994).

The case I have just described is more complex than the case of simple intentional ascent. Suppose I know (and have not momentarily forgotten) that Romeo and Juliet is set in Verona. This is a first-order intentional fact about me. I now execute a simple intentional ascent: as a result, I now know that I know that Romeo and Juliet is set in Verona. In simple ascent, the content of the first-order representation is replicated within the content of the second-order state, along with a representation of my first-order psychological attitude towards that content. The content of my knowing that p is *p*. The content of my knowing that I know that p is *my knowing that p*. The content of the lower-order state, p, is replicated within the content of the higher-order state. There are two distinct cognitive states here, both of which include the content p, but one of which also includes a representation of my lower-order attitude to p. That is why, representation of psychological attitudes is essential for intentional ascent; otherwise, there is no distinguishing intentional ascent from simple conjunction of first-order psychological states with a common content, as when I both believe that p and also regret that p. Replication of content is an important fact about metacognition when it involves simple intentional ascent. It is an important fact to consider when we are thinking about the functions of metacognition. If I know that I know that p, what new capacities do I get that I lack if I simply know that p? How does metacognition enhance our basic, first-order cognitive capacities? Behavioural tests for metacognition aim to detect exactly those enhancements. The tests are difficult to devise in proportion to the difficulty of determining exactly what cognitive enhancements are produced by ascent to the metacognitive level, let alone finding behavioural evidence for the presence of metacognitive states.

The feeling-of-knowing cases are more complex than the simple cases of ascent. If it is true that I know the setting for Romeo and Juliet, then I know that Romeo and Juliet is set in Verona. But if this is a FOK-case, then the fact that the setting is *Verona* is not represented at the higher level. The lower-order content is incompletely represented at the higher-order. The lower-order state would actually be more useful to me than the corresponding higher-order state – except of course that the lower-order state is currently inaccessible to me, is not actively present to my consciousness.

Uncertainty in dolphins

Suppose I am in a state of cognitive uncertainty. Suppose that I am trying to determine whether one perceptual stimulus, p, is the same as or different from another stimulus, q; but because they are very similar, I cannot quite decide, one way or the other. I am uncertain. Uncertainty is a first-order cognitive state. If I also know that I am uncertain, then I am in a further, second-order cognitive state. The central thesis of the research into metacognition in dolphins is that dolphins can know of their own states of uncertainty (Smith et al. 1995, 2003).

The dolphin was trained in a perceptual discrimination task. He learned first to identify a particular stimulus: a sound at a fixed frequency. This is the sample sound. Later on, he learned to match other sounds to his stored template of the sample sound. He was presented with a sound which was either the same in pitch as the sample sound or different in pitch. He had to respond in one way if it was the same pitch, another way if it was a different pitch. The sample sound was not separately available during these trials: the dolphin was forced to rely on memory.

The trained dolphin is now introduced to the test environment. The tests aim to create cognitive difficulties for the dolphin. They force him to make difficult discriminations at his perceptual threshold. The intent is to create in him a state of uncertainty. The question then becomes whether he can use his own uncertainty to produce a specific response. To test for the capacity to take advantage of his own uncertainty, the dolphin is provided with a response that is rewarded if but only if he is uncertain. Here is a useful analogy for thinking about this test environment (Smith et al. 2003): 'imagine that traffic lights gradually morphed from red to green, and that drivers decided whether their light was green enough to go'. The problem of decision faced by the dolphin is like the problem of decision faced by drivers in such a world.

In more detail: the subject in the experiments was an 18-year-old bottlenose dolphin named Natua. He was tested on an auditory discrimination task. The initial task is to learn to recognize tones that sound at a frequency of 2100-Hz (the High tone). The dolphin then learns to discriminate the High tone from tones of lower frequency (Low tones). Training begins with easy discriminations: early sessions contain only High tones at 2100-Hz and Low tones at 1200-Hz. The dolphin learns to press the left (High) paddle for a High tone and the right (Low) paddle for a Low tone. Then the frequency of the Low tone is increased. The dolphin's discrimination threshold at 2100-Hz is about 15-Hz: there is for him a just-noticeable difference between 2085 and 2100-Hz. We can predict then that the dolphin's performance will fall to chance levels, somewhere within this 15-Hz range. It is true that within this range, he will sometimes classify a stimulus correctly. However, given the psychophysical data about his discrimination thresholds, we must infer that in such cases, the success is a matter of chance: a lucky guess. Behavioural success in these cases is not evidence for a cognitive achievement, the animal's recognition that a

stimulus belongs to a specific stimulus class. The animal cannot recognize this because he lacks the requisite discriminatory capacity.

As expected, the test results show that the trained dolphin performs well, with a low error rate, on low frequency tones. His error rate climbs rapidly as the tones approach his discrimination threshold. Even at 2100-Hz, when the High response is required, the dolphin presses the Low paddle 20% of the time. His earlier, highly trained ability to recognize High tones as High has now been compromised, as a result of his exposure to a range of difficult tones that are hard for him to classify.

When the dolphin recognizes a High tone, he presses the High paddle. When he recognizes a Low tone, he presses the Low paddle. In the region of his discrimination threshold, he will have difficulty deciding whether the tone is High or Low. But no specific response has yet been provided for him when he is in this state of uncertainty or indecision. At the next stage of the experiment, a third paddle is introduced: the Escape paddle. The reward schedule is constructed to make the choice of this paddle equivalent to declining the trial. The Escape response itself is not directly rewarded or penalized. When the dolphin is not able, with any confidence, to assign the stimulus to either the High class or the Low class, he can decline to press either of those paddles and press the Escape paddle instead. His choice of the Escape paddle in this situation is interpreted by the authors of the study as evidence that he is aware of his own state of uncertainty: he has some knowledge of his own state of mind.

The rationale for this discrimination threshold experiment is that higher-level cognitive activity often occurs in response to the unusual or the challenging. A high level of difficulty in a task can be used to induce higher levels of cognitive processing in the subject. In routine situations, animals rely on practised responses that are under the control of lower-level cognitive systems. In unusually challenging circumstances, animals are forced to give of their cognitive best. This is where we should be looking for evidence of reasoning, deliberation, consciousness, metacognition, and other components of advanced mental activity. In the case at hand, the difficulties for the dolphin are caused by setting a difficult discrimination task. The Escape response is then interpreted as evidence of a metacognitive state, a state in which he is aware of the difficulty he is having in completing the task. The dolphin is unable to assign a particular stimulus to either of two available stimulus classes. This induces a state of indecision. This indecision is reported to higher cognitive systems in the dolphin's mind. Meta-level cognition occurs. He comes to know that he cannot tell whether the tone is High or Low. So he chooses instead the Escape response.

The argument for metacognition in dolphins

The argument for the metacognitive interpretation of the dolphin's choice at threshold of the Escape response depends on three main points:

(1) Similar response patterns can be elicited from other species, including humans.
(2) Humans report that their choice of the Escape paddle is motivated by their state of uncertainty about the stimulus: they recognize their own uncertainty and choose Escape for that reason.
(3) It is parsimonious to explain close similarities in complex behaviour patterns, both within and especially between species, by invoking the same kind of psychological mechanism.

I now comment on each of these three points in turn. (1) Uncertainty monitoring experiments have been conducted with a variety of other test subjects, including humans, rhesus monkeys and rats. The response patterns of human subjects are similar to the response patterns of the dolphin. One difference is that humans employ the Escape response much less often than do dolphins: 20% at threshold, against 45%. As the authors comment, this probably has more to do with human bravado than with any deep cognitive difference. In another experiment in which humans and rhesus monkeys were both tested on a visual discrimination task, the performance graphs were very similar indeed, representing (say the authors) some of the strongest known parallels between human and animal performance on such tasks (Smith et al. 2003).

There is another strand in the pattern of similarities between human and animal subjects. A well-known phenomenon in human metacognition research is over-confidence. When human subjects are asked to report on the probability that they have remembered a fact correctly, or have found the correct solution to a problem, or that the name they have on the tip of their tongue is the right name, they typically report probabilities of being correct that are too high. Actual success rates are consistently lower than subjects' own expected success rates. Over-confidence is a robust feature of human cognitive psychology. Amazingly, dolphins are similarly over-confident (Smith et al. 1995). When their actual performance is compared with a simulation that uses optimal criteria, it turns out that dolphins, like humans, do not make enough use of the Escape response. They have too much confidence in their own discriminatory capacities. Like humans, they would improve their overall rewards by declining challenges more often.

These similarities in the response profiles of human and animal subjects give us the first premiss of an argument for the conclusion that the dolphin's choice of the Escape response is prompted by a metacognitive state, the knowledge of his own uncertainty.

(2) When human subjects are asked to report on their mental processes in this test situation, they distinguish between the state of mind that leads to High or Low responses, on the one hand, and the state of mind that leads to the Escape response, on the other hand. The High and Low responses are produced, they say, in response to properties of the stimulus: the distinctive sounds of High and Low tones. But when the stimulus is hard to classify, subjects

report that the determining factor is their own state of doubt and uncertainty (Smith et al. 1995). In this perceptual condition, subjects feel uncertain or doubtful about the correct assignment of the stimulus. They choose the Escape response because they know that they do not know whether the stimulus is High or Low. This information about the psychological mechanisms that are responsible for choice of the Escape response in human subjects gives us the second premiss of the argument.

(3) Rhesus monkeys and human subjects produce almost identical response patterns in a difficult discrimination task. Human subjects consistently provide a metacognitive explanation for their Escape responses. It is parsimonious to suppose that monkeys use the same kind of psychological mechanism to produce their own closely corresponding patterns of Escape responses. Same behavioural effects; therefore, probably, same psychological causes. A claim about explanatory parsimony in comparative psychology takes us to the conclusion that a metacognitive interpretation for the dolphin's behaviour is confirmed.

However, is phylogenetic distance a confounding factor here? Rhesus monkeys are phylogenetically a lot closer to us than are dolphins. Rats, at a greater phylogenetic distance from us than monkeys, make very infrequent selections of the Escape response at their perceptual thresholds, and show no improvement with extended training. Dolphins are more distant to us than are rats. To what extent can behavioural similarities between humans and dolphins be used as evidence for psychological similarities? The argument would fail if it referred to cognitive mechanisms that were peculiarly primate. It would fail if uncertainty monitoring were just the primate way of doing a job that could be done in other ways. But metacognition is not like that. There are generic ways of building lower-order cognitive systems (the most basic being belief-desire psychology). Take such a system and add a capacity for metarepresentation. The result is a system with generic metacognitive capacities. There is a plausible case, along these lines, for the view that the comparative argument from humans (and other primates) to dolphins rests on principles of universal psychology, not on species-specific principles of primate psychology.

Two other considerations strengthen the case for the metacognitive interpretation of the dolphin's Escape response. The first is that the perceptual threshold results are not isolated results. Smith and his collaborators describe a memory monitoring task that has been tested on a variety of species. Some species succeed, some fail, in a pattern that roughly corresponds with species-specific successes and failures on the perceptual threshold tasks. Success on the memory monitoring tasks requires the use of an 'indeterminacy resolution' process. A parsimonious explanation for success on both the perceptual threshold task and the memory task is that the animal is using metacognitive capacities, of the same general kind, in both cases. If metacognitive activity is ruled out, there is a risk of loss of explanatory generality. There is loss of generality if different kinds of lower-level explanations are adopted for the two tasks. But it is unparsimonious to adopt one kind of lower-level explanation

for the animal's responses on one task and a different kind of lower-level explanation for the animal's responses on the other task, when a single kind of metacognitive explanation is available for both tasks. Parsimony does not always count in favour of explanations at 'lower levels of the psychological scale'.

The second supporting consideration rests on the relationship between High and Low responses, on the one side, and Escape responses, on the other. The presence of the Escape paddle gives the dolphin three choices of response, where previously he had two. There is evidence (discussed shortly) that the choice of High and Low responses is mediated by cognitive processes, albeit first-order cognitive processes. If this interpretation is accepted, then psychological plausibility suggests that the selection of the third, Escape response is mediated by processes that are, at a minimum, no less complex. If the choice between the High and Low responses is mediated by a judgement that the stimulus is the same as or different from the sample stimulus, then the Escape response is probably mediated by a cognitive process that is yet more complex. What is not settled is whether this more complex process involves meta-level ascent.

Deflationary hypotheses

Is there an alternative to the metacognitive hypothesis? In particular, is there a deflationary alternative, an interpretation that does not attribute metacognitive powers to the dolphin?

Smith and his co-workers describe one alternative hypothesis which might appear to explain the dolphin's Escape response. But the hypothesis they consider and reject is not a first-order cognitive hypothesis, not even a cognitive hypothesis at all. It is simply that the dolphin's Escape response is under stimulus control (Smith et al. 2003). Stimulus control is a behaviourist concept. An action is under stimulus control if the stimulus event leads, more or less automatically, to the behavioural response.

If a pattern of behaviour is under stimulus control, then there must be a matching stimulus class: that is, there must be a *kind* of stimulus to which the behaviour becomes attached. But (the argument goes) there is no stimulus class which is intermediate between the High stimulus class and the Low stimulus class. There are just those two stimulus classes, and a region adjacent to the boundary between them in which the subject has discrimination problems. If there is no distinct stimulus class corresponding to the Escape response, then that response cannot be under stimulus control.

There is a supplementary argument which, in my view, tells strongly against noncognitive interpretations of the Escape response. It is an argument from comparative psychology. Rats do not perform on the discrimination task as do humans, monkeys and dolphins (Smith et al. 2003). Rats perform competently on the High and Low response tasks, but fail to learn the Escape response in

the region of their perceptual threshold. Yet rats are perfectly capable of learning to give a unique response to a middle stimulus class. If there really were High, Low and Middle stimulus classes in this situation, then rats should be able to learn a unique response for each stimulus class. But rats do not learn to select Escape in response to tones adjacent to their discrimination thresholds. There is no usable stimulus class corresponding to the Escape response. Take the set of (objective) tones that cause perceptual difficulty: there is no detectable feature of those tones onto which the rats can perceptually latch and which they can then use to assign those tones to an intermediate class. The argument from comparative psychology generalizes this result to other species. If there is no stimulus class corresponding to a response class, and some species (humans, monkeys and dolphins) do exhibit the Escape response, then we should conclude that in those species, the Escape response is not under stimulus control. Those species must be employing further levels of cognitive processing, levels that the rat either does not access or cannot access.

This comparative argument makes a case for the conclusion that dolphins can use levels of cognitive processing beyond the level of sensory processing when they construct their responses to perceptual difficulty. This is an independently important result. It indicates that only a creature capable of first-order cognition can produce the pattern of results exhibited by the dolphin in the uncertainty monitoring experiments. Persuasive evidence for first-order cognition is always welcome. But the argument does not show that the further processing involved in the selection of the Escape response extends above the level of first-order cognition into metacognition. Complex cognitive processing which remains at the level of first-order representation must not be mistaken for processing at a higher order.

Consider an easy discrimination task: the dolphin learns to press the Low paddle when the tone sounds at 1200-Hz and the High paddle when the tone sounds at 2100-Hz. It is unlikely that these responses are under stimulus control. Smith agrees (personal communication) that a cognitive explanation should be provided for the dolphin's selection of High and Low responses in easy discrimination settings (prior to his introduction to confounding threshold tones). If this is correct, then a High or Low tone causes the dolphin to move deliberately and purposefully towards the appropriate paddle, with his behaviour giving little if any evidence of being an automatic reflex. In such a case, the stimulus of hearing the tone would not directly cause a behavioural response. It would give rise instead to a cognitive state, and the dolphin's subsequent behaviour would be under the control of this cognitive state. We know nothing about the representational systems that dolphins use. So I will simply suppose that the dolphin's cognitive response to a High or a Low tone is a perceptual judgement, a judgement that the tone is High or Low. These are first-order cognitive states.

The cognitive interpretation of the High and Low responses strengthens the case for a metacognitive interpretation of the Escape response. As noted above, psychological realism suggests that the Escape response is not less complex

than the High and Low responses. Escape is chosen only when the animal has tried but failed to reach a same/different judgement.

It is a reasonable conclusion that the cause of the Escape response is not less complex than the causes of the High and Low responses. But this is consistent, in principle, with a first-order cognitive explanation of the Escape response. Perhaps the dolphin recognizes the tone as being neither High nor Low but as having some third quality. This is still a first-order cognitive hypothesis. The problem is to find a suitable third quality that can be recognized in the tone. A tone that falls near the threshold has no special perceptual characteristics that mark it out from High and Low tones. (It is not blurred, or lacking a definite pitch). As the argument from comparative psychology shows, the sensory state which is produced in rats by that kind of stimulus does not contain information to distinguish it from High and Low tones, information that could be harnessed to drive an Escape response.

Uncertainty

Smith and his collaborators describe the dolphin's first-order state, in the zone of perceptual difficulty, as a state of uncertainty. A comparative argument is the justification for the choice of this concept: uncertainty features in human post-experimental reports. Yet uncertainty is a concept that carries a lot of philosophical baggage. Descartes recommended that we respond to uncertainty by suspending belief. My view is that this model of uncertainty as the suspension of belief plays an important but unacknowledged role in the argument for the metacognitive interpretation of the dolphin's behaviour. On the 'Cartesian' interpretation, the dolphin's psychological state consists of a sensory state (hearing a tone) plus a cognitive state of suspended belief (suspending belief about whether the tone is High or Low). When looking for alternatives to a metacognitive interpretation of behaviour, the first place to look is to the level of first-order cognitive states. However, in the present case, the only relevant first-order cognitive state is a state of uncertainty, a state of suspended belief. The first-order cognitive state of the Cartesian sceptic is not one of belief but of the suspension of belief, not a state of knowing but a state of not knowing. The state of not knowing, however, has little if any usable representational content. Suspension of belief has more in common with the absence of cognition than its presence.

One can escape from this cognitive vacuum by going metacognitive. One ascends from not knowing to knowing that one does not know, ascends from the absence of knowledge to knowledge of the absence of knowledge. Knowing that one does not know is a state that does have some useful representational content. It represents the fact that one does not know enough to justify the choice of either the High or the Low response. This kind of psychological self-knowledge can explain the dolphin's choice of the Escape response in the discrimination tasks.

First-order hypotheses

Is there a deflationary alternative that remains at the level of first-order cognition? That depends on our finding a suitable first-order cognitive state that does have usable content. We need to find another way to describe the first-order cognitive state that the dolphin is in at his perceptual threshold. We need an alternative to the idea that when he is in perceptual difficulty, he is in a state of suspended belief. Here is such an alternative.

The new hypothesis is that the dolphin is not in a state of cognitive suspension but one of cognitive conflict. It is not the case that he fails to identify the tone as either High or Low, but that he identifies it as both. There are various ways of describing this state of cognitive conflict. Perhaps his state is a simple conjunction of the perceptual judgement that the tone is High and the perceptual judgement that the tone is Low. Logical consistency among propositional attitudes is never a psychological necessity. Or perhaps the dolphin assigns a high probability to the hypothesis that the tone is High, a probability sufficient for acceptance, and assigns the same high probability to the hypothesis that the tone is Low. Or perhaps he oscillates between two cognitive attractors: the belief that the tone is High and the belief that the tone is Low. These are all possible states for the dolphin to be in when he hears the tone at threshold. They are all first-order cognitive states, or better, conjunctions of such cognitive states. In all of the psychological conditions that I have just described, the dolphin's attention is focussed on something nonmental: the sounding tone. His attention is not focussed on something mental: he is not attending to his own first-order cognitive responses to the sounding tone.

To complete this sketch of a hypothesis, we need to sketch a mechanism that allows this state of conflict to be recruited to drive the Escape response. This presents no difficulties. It is true that one possible mechanism is metacognitive: the dolphin recognizes his state of cognitive conflict and chooses to bail out of the trial by taking the Escape response. The lower-order state is causally effective but only indirectly, only because it gives rise to a higher-order state which controls the behavioural response. But there are alternative, lower-order mechanisms. A decision rule, one that is activated when the subject is in a state of cognitive conflict, is one possibility. A rather different possibility is that the state of cognitive conflict brute-causes some noncognitive state that can in turn be recruited to drive the Escape response. This is a rich source of psychological hypotheses. Perhaps being in a state of cognitive conflict feels unpleasant or makes the animal nervous or anxious. In this kind of case, a cognitive condition causes an affective state that causes the animal to choose the Escape response.

In the cases just described, a first-order cognitive state is recruited by some other inner state to drive an adaptive response. The behaviour is caused by a conjunction of psychological states, none of which rise above first-order intentionality. Metacognition is a special kind of way in which (first-order) cognitive states are taken up and exploited by other psychological mechanisms.

But it cannot be the only way in which first-order cognitive states are put to work. Even a first-order state of uncertainty could, in principle, be harnessed by other first-order mechanisms and put to adaptive use. If there are doubts about such an explanation for the dolphin's choice of the Escape paddle, it is because uncertainty, on a Cartesian understanding, has insufficient representational content to be recruited to the task of driving a specific response.

I have sketched several explanations for the dolphin's choice of the Escape response. All of the explanations are cognitive, but only some of them impute metacognition to the dolphin. This situation is ubiquitous in animal psychology research. When several competing hypotheses are on the table, each of which is (we will suppose) a possible candidate for an explanation of some behavioural data, how is it reasonable to proceed? Where do we go next?

Probability, not possibility, is the issue

The dolphin's Escape response can be explained metacognitively. It can also be explained as a first-order cognitive response. Perhaps it is reasonable for us to accept the metacognitive interpretation only if all possible first-order explanations have first been eliminated. But if this were sound methodology, then any argument for animal metacognition could be rebutted simply by inventing a lower-order hypothesis that had not yet been specifically disconfirmed. Similarly, any argument for animal cognition could be rebutted by inventing a noncognitive explanation that had not yet been specifically disconfirmed. Some of the debates about animal cognition have this flavour (e.g. Heyes 1998). But this is a methodology that leads directly to Cartesian scepticism. Cartesian scepticism is produced by the thought that it is reasonable to accept any hypothesis only if all possible alternative hypotheses have been rebutted. But this is not a sound methodology for empirical science. In science, empirical probability is the issue, not logical possibility. The sheer logical possibility that the Escape response has been generated by a (lightly sketched) first-order mechanism is not of itself sufficient reason to withhold acceptance of the metacognitive hypothesis, if the metacognitive hypothesis already has some empirical support. (If the metacognitive hypothesis were simply one more logical possibility, then it would be relevant to point to other logical possibilities).

Smith et al. have made an empirical case for the metacognitive hypothesis. The case is not just that a response could possibly be produced metacognitively. Instead, a complex pattern of responses in dolphins is very similar to a complex pattern of responses in humans (and monkeys), when members of these species are exposed to the same problem-setting environment. There is independent evidence that the human subjects generate this behaviour pattern metacognitively. It is parsimonious to explain similar, complex, stimulus–response patterns by similar psychological mechanisms. Therefore, the evidence currently available makes it reasonable to accept, provisionally, that the

dolphin's response pattern is also produced by metacognitive processes. The introduction of alternative hypotheses for which there is, at present, no specific empirical support is at most an invitation to further research; it is not by itself a rebuttal of the metacognitive hypothesis.

Intentionality and Lloyd Morgan's Canon

These methodological observations do not, however, address the specific concern that is iconically represented as Lloyd Morgan's Canon. I have already suggested that the Canon has at least one plausible application, namely, to cases involving intentional ascent. If an animal's behaviour can be explained in two different ways, and one of those explanations invokes a higher order of intentionality than the other, then if other things are equal, we should prefer the lower-order explanation. The reasons I earlier gave were of a general nature: because of the dependency relations between them, higher-order thoughts must occur less frequently than lower-order thoughts; and higher-order thought is also likely to be more energetically demanding. But a more exact reason can be given when intentional ascent is specifically implicated. In order for there to be higher-order intentionality, there must be representation of other psychological attitudes. Thinking about my belief that p constitutes higher-order thought if the content of the thought includes a representation not only of p but of my believing that p. In turn, the obvious model of this capacity requires my possession of psychological concepts: in this example, the concept of believing. Intentional ascent is possible only if appropriate concepts for intentional attitudes are to hand. This is how I understand the distinction drawn by Andrew Whiten (Whiten 2000). Whiten distinguishes between two different concepts of meta-representation. These are: (1) a mental representation of a mental representation, and (2) a mental representation of a mental representation *as a representation*. To have a concept of the mental, he says, it is not enough just to be able to register and respond to some of one's own mental processes; one needs to be able to register them *as* mental processes (see also Sperber 2000, for a similar line of thought).

With this material to hand, I can now complete the defence of the limited application of Lloyd Morgan's Canon, its application specifically to cases in which there is competition between hypotheses at different levels of intentionality. Higher-order intentionality depends on possession of psychological concepts. Psychological concepts are 'theoretical' concepts, relative to concepts for observable objects and properties in the physical environment. The capacity to understand the world in terms of unobservable processes and structures represents a very significant intellectual advance on the simpler capacity for understanding the world in terms of associations between observables. An explanation which does not presuppose that the animal in question has made that intellectual advance is to be preferred to one that presupposes that it has made the advance, other things being equal. Of course, a consilience of

evidence that the animal has made the intellectual advance can defeat a presumption of this kind.

If there is a first-order intentional explanation for the dolphin's choice of the Escape response, then it should be preferred to any second-order intentional explanation, all else being equal. Consider the hypothesis (H1) that at threshold, the dolphin feels uncertain or conflicted, and that somehow, this feeling is directly recruited to drive the choice of the Escape response. Consider now the hypothesis (H2) that at threshold, the dolphin feels uncertain or conflicted, and that he knows or is aware that he is in this psychological state: he 'knows that he does not know' whether the tone is High or Low. This higher-order state of knowing (or awareness) drives his selection of the Escape response. In this situation, the defensible interpretation of Lloyd Morgan's Canon warrants our choosing H1 over H2, other things being equal. Intentional ascent should be undertaken only if no reasonable alternative is available. But this is not the end of the matter.

The function of metacognition

The psychologist's functional concept of metacognition is neutral on the question of intentional ascent. Monitoring and control processes must be commonplace in physiology. Think for example about a device that monitors muscle tension, receiving information about stresses and sending out appropriate control signals. No conceptual thought is necessary: the device does not apply concepts of stress or relaxation to the muscle groups that it regulates. Similarly, functional metacognition does not absolutely need to be a conceptually mediated activity. In other words, the capacity for making adaptive *cognitive* responses to some of one's own cognitive processes and states does not depend essentially on a capacity for intentional ascent. If this is correct, then the prohibition on unnecessary intentional ascent does not rule out some metacognitive interpretations of the dolphin's Escape response. Perhaps 'uncertainty' is not the best concept to apply to the cognitive state induced in the dolphin at threshold. Perhaps 'knowing that one does not know' is too redolent of intentional ascent and competence with psychological concepts to be the best model for the dolphin's metacognitive activities. We do not know what first-order state the dolphin is in when he hears tones at threshold which he cannot easily classify. Perhaps this initial state is indecision, perhaps it is conflict. In either case, a metacognitive hypothesis is available that does not imply intentional ascent. The initial state is detected by another cognitive system, one which is alerted to a problem that has arisen in the dolphin's ordinary cognitive endeavours. It is not, perhaps, the initial state by itself that triggers a warning, but what it implies in the context of a wider endeavour: the match-to-sample task. The detector state, responding to uncertainty or conflict or whatever, is functionally a metacognitive state. And like any other cognitive state, it can (in principle) be recruited to drive the Escape response.

The functional concept of metacognition does not specify that the output of the monitoring device is an intentional state. Only if psychological self-monitoring produces intentional states does it implicate higher-order intentionality: a thought about the monitored state. In the basic functional model, what monitoring produces is a 'partial model' of the state being monitored. This much at least is necessary, if this concept of metacognition is not to collapse back into the concept of further, first-order processing. But there is no particular reason to think that the partial model must include a concept for the kind of first-order psychological state that is being monitored. It is enough that the lower-order state is discriminated from some other possible states that might instead be present, such as the states of recognizing the tone as High and recognizing it as Low.

Indeed, the principle of parsimony can be preserved by interpreting the human responses in the same way. Human subjects monitor their own cognitive performance on the threshold problems, discriminate the state they are in as different from both the state of recognizing the tone as High and the state of recognizing it as Low, and consequently produce a third response. When the post-experimental probe is applied, then (and only then) do they undertake an intentional ascent and apply psychological concepts to their own earlier performance. The probe, not the experiment, induces intentional ascent.

On deflationary conclusions

The functional concept of metacognition does not entail intentional ascent. Intentional ascent occurs only if the output of the psychological self-monitoring device is interpreted under psychological concepts. I have suggested that monitoring devices could perform useful functions without generating propositional attitudes as outputs. The biggest obstacle to acceptance of the hypothesis of dolphin metacognition is presumably the assumption that dolphins must in that case possess concepts for psychological states. This is an assumption about the kinds of states that these monitoring devices produce. But if psychological self-monitoring need not produce propositional attitudes as output, then the hypothesis of dolphin metacognition survives this source of criticism.

Does this interpretation of dolphin metacognition counts as a deflationary interpretation, in the wider context of research into 'higher' mental processes in nonhuman animals? A capacity for mental activity that is functionally metacognitive, and yet which does not involve intentional ascent, may seem to have only modest significance. It is certainly less than the capacity for thinking about intentional interpretations of behaviour. The capacity to adopt 'the intentional stance' is the specific capacity that is alleged, with some justification, to be a necessary condition for some of the most distinctive mental and cultural achievements of human beings: language, science, art, politics, and so on. Perhaps the best explanation for the behavioural data in the perceptual

discrimination tasks is that the dolphin has the capacity to register that he is in a state of cognitive difficulty, a capacity that does not require the application to himself of specifically psychological (intentional) concepts. He is able to recruit this rather modest kind of self-understanding to drive adaptive behaviour. This is not by itself evidence that the dolphin has the capacity to attribute intentional states to others or to himself. For this reason, the research on 'uncertainty monitoring' in dolphins, monkeys, and other species, is not closely connected to the primate research on theory of mind. The latter, but not the former, is (it seems) committed to intentional ascent.

Metacognition and consciousness

These conclusions constitute the more negative outcome of my inquiry. But there is a more positive outcome as well. I have argued that cognition about cognition can occur in the absence of competence in the use of intentional psychological concepts. What kinds of interesting psychological capacities might be entrained by this kind of metacognition? One obvious possibility is that functional metacognition is the basis for one form of psychological self-awareness, one of the varieties of self-consciousness. William Lycan, among others, has defended a metacognitive theory of consciousness as the product of psychological self-monitoring processes (Lycan 1987). If self-monitoring is understood on the functional model, then it does not necessarily entrain intentional ascent and the consequent possession of psychological concepts. In that case, even a comparatively restrained, modest and noninflationary interpretation of the dolphin's performance on these 'uncertainty monitoring' tasks, an interpretation of the kind described above, might provide a reason for attributing to the dolphin a suitably modest, but real, variety of self-consciousness.

References

Carruthers P. and Smith P. (eds). 1996. Theories of Theories of Mind. Cambridge University Press, Cambridge.

Dennett D. 1987. Intentional systems in cognitive ethology. The Intentional Stance. MIT Press, Cambridge, MA.

Herman L. 2002. Exploring the cognitive world of the bottlenosed dolphin. In: Bekoff M., Allen C. and Burghardt G. (eds), The Cognitive Animal. Empirical and Theoretical Perspectives on Animal Cognition. MIT Press, Cambridge, MA.

Heyes C. 1998. Theory of mind in nonhuman primates. The Behavioral and Brain Sciences 21: 101–48.

Lycan W. 1987. Consciousness. MIT Press, Cambridge, MA.

Metcalfe J. and Shimamura A. (eds). 1994. Metacognition. Knowing About Knowing. MIT Press, Cambridge, MA.

Morgan C. 1894. An Introduction to Comparative Psychology. Walter Scott, London.

Nelson T. 1996. Consciousness and metacognition. Am. Psychol. 51: 102–116.

Smith J.D., Schull J., Strote J., McGee K., Egnor R. and Erb L. 1995. The uncertain response in the bottlenosed dolphin (Tursiops truncatus). J. Exp. Psychol. Gen. 124: 391–408.
Smith J., Shields W. and Washburn D. 2003. The comparative psychology of uncertainty monitoring and metacognition. Behav. Brain Sci. 26: 317–373.
Sober E. 1998. Morgan's canon. In: Cummins D. and Allen C. (eds), The Evolution of Mind. Oxford University Press, New York.
Sperber D. 2000. Metarepresentations in an evolutionary perspective. In: Sperber D. (ed.), Metarepresentations: A Multidisciplinary Perspective. Oxford University Press, Oxford.
Tomasello M. 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge MA.
Whiten A. 2000. Chimpanzee cognition and the question of mental re-representation. In: Sperber D. (ed.), Metarepresentations: A Multidisciplinary Perspective. Oxford University Press, Oxford.
Whiten A. and Byrne R. 1988. Tactical deception in primates. Behav. Brain Sci. 11: 233–73.