

Comparative Approaches to Cognitive Science

edited by Herbert L. Roitblat and Jean-Arcady Meyer

on, advisors

Analysis with
H. Holland

First European
Paul Bourguine
ans of Natural

nal Conference
eyer, Herbert

and Thomas

near, Jr.

ograms, John

Parallel Micro-

nal Conference
lip Husbands,

orkshop on the
A. Brooks and

ert L. Roitblat

ior, J. A. Scott

1995

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

LIBRARY
UNIVERSITY OF NORTHERN IOWA
CEDAR FALLS, IOWA

6

Do Animals Have Beliefs?

Daniel C. Dennett

According to one more or less standard mythology, behaviorism, the ideology and methodology that reigned in experimental psychology for most of the century, has been overthrown by a new ideology and methodology: cognitivism. Behaviorists, one is told, did not take the mind seriously. They ignored—or even denied the existence of—mental states such as beliefs and desires, and mental processes such as imagination and reasoning; behaviorists concentrated exclusively on external, publicly observable behavior and the (external, publicly observable) conditions under which such behavior was elicited. Cognitivists, in contrast, take the mind seriously and develop theories, models, and explanations that invoke, as real items, these internal, mental goings-on. People (and at least some other animals) have minds after all; they are *rational agents*.

Like behaviorists, cognitivists believe that the purely physical brain controls all behavior without any help from poltergeists or egos or souls, so what does this supposedly big difference come to? When you ask a behaviorist what the mind is, the behaviorist retorts, "What mind?" When you ask a cognitivist, the reply is, "The mind is the brain." Since both agree that it is the brain that does all the work, their disagreement looks at the outset to be merely terminological. When, if ever, is it right, or just perspicuous, to describe an animal's brain processes as thinking, deciding, remembering, imagining? This question suggests to some that the behaviorists may have been right about lower animals—perhaps about pigeons and rats, and certainly about frogs and snails; these simple brains are capable of nothing that should be dignified as properly "cognitive." Well, then, where do we "draw the line" and why?

Do animals have beliefs? One of the problems with this question, which has provoked a lot of controversy among animal researchers and the ideologues of cognitive science, is that there is scant agreement on the meaning of the term "belief" as it appears in the question. "Belief" has come to have a special, nonordinary, sense in the English of many (but not all) of these combatants: it is supposed by them to be the generic, least-marked term for a cognitive state. Thus, if you look out the window and *see* that a cow is in the garden, you *ipso facto* have a belief that a

cow is in the garden. If you are not ignorant of arithmetic, you believe the proposition that $2 + 2 = 4$ (and an infinity of its kin). If you *expect* (on whatever grounds) that the door you are about to open will yield easily to your tug, then you have a belief to that effect, and so on. It would be more natural, surely, to say of such a person, "He thinks the door is unlocked" or "He is under the impression that the door is open" or, even less positively, "He does not know the door is locked." "Belief" is ordinarily reserved for more dignified contents, such as religious belief, political belief, or—sliding back to more quotidian issues—specific conjectures or hypotheses considered. But for Anglophone philosophers of mind in particular, and other theoreticians in cognitive science, the verb "believe" and the noun "belief" have been adopted to cover all such cases; whatever information guides an agent's actions is counted under the rubric of belief.

This particularly causes confusion, I have learned, among non-native speakers of English; the French term "*croissance*," for instance, stands even further in the direction of "creed" or "tenet," so that the vision my title question tends to conjure up for Francophones is an almost comical surmise about the religious and theoretical convictions of animals—not, as it was meant to be understood, a relatively bland question about the nature of the cognitive states that suffice to account for the perceptuolocomotory prowess of animals. But even those Anglophones who are most comfortable with the artificially enlarged meaning of the term in their debates suffer, I think, from the same confusion. There is much less agreement than these theorists imagine about just what one would be asserting in claiming, for instance, that dogs have beliefs.

Consider the diversity of opinion. Do animals have beliefs? I have said yes, supporting my claim by pointing to the undeniable fact that animals' behavior can often be predicted (and explained and manipulated) using what I call the intentional stance (Dennett 1971, 1987)—the strategy of treating animals as rational agents whose actions are those they deem most likely to further their "desires" given their "beliefs." One can often predict or explain what an animal will do by simply noticing what it notices and figuring out what it wants. The raccoon wants the food in the box-trap, but knows better than to walk into a potential trap where it cannot see its way out. That is why you have to put two open doors on the trap—so that the animal will dare to enter the first, planning to leave by the second if there is any trouble. You will have a hard time getting a raccoon to enter a trap that does not have an apparent "emergency exit" that closes along with the entrance.

I take it that this style of explanation and prediction is uncontroversially valuable: it works, and it works *because* raccoons (for instance) are that smart. That fact suffices, given what I mean by "belief," to show that raccoons have beliefs—and desires, of course. One might call the latter preferences or goals or wants or values, but whatever you call them, their

specification involves the use of intentional (mentalist) idioms. This guarantees that translating between "desire" talk and "preference" or "goal" talk is trivial, so I view the connotational differences between these terms as theoretically irrelevant. The same thing holds for beliefs, of course: you might as well call the state of the raccoon a belief, since if you call it a "registration" or a "data-structure" in the "environmental information store" or some other technical term, the logic you use to draw inferences about the animal's behavior, given its internal states, will be the standard, "intentionalist" logic of belief. For more on the logic of intentionality, see Dennett (1969, 1971, 1983, 1987) or the article on intentionality in the *Oxford Companion to the Mind* (Gregory 1987).

When called upon to defend this indifference to terminological niceties, I like to point out that when economists, for example, consider the class of *purchases* and note the defining condition that the purchaser believes he is exchanging his money for something belonging to the seller and desires that item more than the money he exchanges for it, the economist is not requiring that the purchaser engage in any particularly salient act of creed endorsing (let alone suffer any spasms of *desire*). A purchaser can meet the defining "belief-desire" conditions while daydreaming, while concentrating on some other topic, while treating the seller almost as if he/it were a post stuck in the ground. All that has to be the case is that the purchaser has somehow or other come into a cognitive state that identifies a seller, a price, and an opportunity to exchange and has tipped the balance in favor of completing the transaction. This is not nothing; it would be a decidedly nontrivial task to design a robot that could distinguish an apple seller from an apple tree while not becoming a money pump when confronted by eager salesmen. But if you succeeded in making a successful purchaser-robot, you would ipso facto have made a robot believer, a robot desirer, because belief and desire, in this maximally bland (but maximally useful!) sense are logical requirements of purchasing behavior.

Others do not approve of this way with words. Donald Davidson (1975), for instance, has claimed that only creatures with the concepts of truth and falsehood can properly be said to have beliefs and, since these are metalinguistic concepts (I am simplifying his argument somewhat), only language-using animals such as human beings can have beliefs. And then there are those who have some other criterion for belief, according to which some animals do have beliefs and others do not. This criterion must be an empirical question for them, presumably, but which empirical question it is—which facts would settle it one way or the other—is something about which there is little agreement. David Premack (1988) has claimed that chimpanzees—and perhaps only chimpanzees—demonstrate belief, while Jerry Fodor (1990) has suggested that frogs—but not paramecia—have beliefs. Janet Halperin (at the conference that resulted in this book) expressed mixed feelings about the hypothesis that her

Siamese fighting fish have beliefs; on the one hand, they do seem richly amenable (in some regards) to intentional interpretation, while on the other hand she has a neural net-like model of their control systems that seems to lack any components with the features beliefs are often supposed to have.

The various assumptions tacitly made about how to use these words infects other controversies as well. Does it follow from the hypothesis that there is something it is like to be a bat (Nagel 1974) that bats have beliefs? Well, could it be the case that there is indeed something it is like to be a bat, but no bat knows what it is like? But could the bat know what it is like without having any beliefs about what it is like? If knowledge entails belief, as philosophical tradition declares, then a bat must have beliefs about what it is like to be it—if it is like anything at all to be a bat. But philosophers have different intuitions about how to answer all these questions, so of course they also have clashing opinions on whether robots could have beliefs.

The maximal leniency of the position I have recommended on this score is notoriously illustrated by my avowal that even lowly thermostats have beliefs. John McCarthy (1979) has joined me in this provocative stance, and he proposes just the right analogy in defense, I think. Is zero a number? Some people were outraged when the recommendation was first made that zero be considered a number in good standing. What kind of a number is zero? It stands for no quantity at all! But the number system you get if you include zero is vastly more perspicuous and elegant than the number system you get if you exclude zero. A thermostat, McCarthy and I claim, is one of the simplest, most rudimentary, least interesting systems that should be included in the class of believers—the class of intentional systems, to use my term. Why? Because it has a rudimentary goal or desire (which is set, dictatorially, by the thermostat's owner, of course), which it acts on appropriately whenever it believes (thanks to a sensor of one sort or another) that its desire is unfulfilled. Of course, you don't have to describe a thermostat in these terms. You can describe it in mechanical terms, or even molecular terms. But what is theoretically interesting is that if you want to describe the set of all thermostats (cf. the set of all purchasers) you have to rise to this intentional level. Any particular purchaser can also be described at the molecular level, but what purchasers—or thermostats—all have in common is a systemic property that is captured only at a level that invokes belief talk and desire talk (or their less colorful but equally intentional alternatives—semantic information talk and goal registration talk, for instance).

It is an open empirical question which other things, natural and artificial, fall into this class. Do trees? The case can be made—and in fact was made (or at least discussed in the appropriate terms) by Colin Allen in the symposium. One can see why various opponents of this view have branded it as “instrumentalism” or “behaviorism” or “eliminative materi-

alism." But before accepting any of these dismissive labels, we should look at the suggested alternative, which is generally called "realism" because it takes seriously the questions Which animals *really* have beliefs and, of those that do, what do they *really* believe? Jerry Fodor (1990), John Searle (1992), and Thomas Nagel (1986) are three prominent philosophical realists. The idea that it makes sense to ask these questions (and expect that, in principle, they have answers) depends on a profound difference of vision or imagination between these thinkers and those who see things my way. The difference is clearest in the case of Fodor, as we can see by contrasting two pairs of propositions:

1. Fodor: *Beliefs are like sentences*. Beliefs have structure, are composed of parts, and take up room in some spatial or temporal medium. Any finite system can contain only a finite number of beliefs. When one claims that Jones believes *that the man in the blue suit is the murderer*, this is true if and only if the belief in Jones's head really is composed of parts that mean just what the words in the italicized phrase mean, organized in a structure that has the same syntactic—and semantic—analysis as that string of words.

1A. Dennett: *Beliefs are like dollars*. Dollars are abstract (unlike dollar bills, which are concrete). The system of dollars is just one of many possible systems for keeping track of economic value. Its units do not line up "naturally" with any salient differences in the economic value of goods and services in the world, nor are all questions of intersystemic translation guaranteed to be well founded. How many U.S. dollars (as of July 4, 1994) did a live goat cost in Beijing on that date? One has to operationalize a few loose ends to make the question meaningful: Do you take the exchange rate from the black market or use the official rate, for instance? Which should you use, and why? Once these loose ends are acknowledged and tied off, this question about the dollar value of goats in Beijing has a relatively satisfactory answer. That is, the various answers that might be reasonably defended tend to cluster in a smallish area about which disagreement might well be dismissed as trivial. How many U.S. dollars (as of July 4, 1994) was a live goat worth in ancient Athens? Here any answer you might give would have to be surrounded by layers of defense and explanation.

Now, no one doubts that a live goat really had value in ancient Athens, and no one doubts that dollars are a perfectly general, systematic system for measuring economic value, but I do not suppose anyone would ask, after listening to two inconclusive rival proposals about how to fix the amount in dollars, "Yes, but how many dollars did it *really* cost back then?" There may be good grounds for preferring one rival set of auxiliary assumptions to another (intuitively, one that pegs ancient dollars to the price per ounce of gold then and now is of less interest than one that pegs ancient dollars to assumptions about "standard of living," the cost per year of feeding and clothing a family of four, etc.), but that does not

imply that there must be some one translation scheme that "discovers the truth." Similarly, when one proposes and defends a particular scheme for expressing the contents of some agent's beliefs via a set of English sentences, the question of whether these sentences—supposing their meaning is fixed somehow—describe what the agent *really* believes betrays a certain naivete about what a belief might be.

2. Fodor: *Beliefs are independent, salient states.*

2A. Dennett: *There are independent, salient states that belief talk "measures" to a first approximation.*

What is the difference between these two propositions? We both agree that a brain filled with sawdust or jello could not sustain beliefs. There has to be structure; there have to be elements of plasticity that can go into different states and thereby secure one revision or another of the contents of the agent's beliefs. Moreover, these plastic elements have to be to some considerable extent independently adjustable to account for the productivity (or, less grandly, the versatility) of beliefs in any believer of any interest (of greater interest than the thermostat).

The difference is that Fodor stipulates that the ascribing language (the sentences of English or French, for instance) must have much the same degrees of freedom, the same planes of revision, the same joints, as the system the sentences describe. I disagree. Consider the information contained in a map drawn on some plane surface according to some mapping rules and utilizing some finite set of labeling conventions. Imagine a robot that locates itself by means of such a system, moving a symbol for itself on its own map as it moves through the world. At any moment, its system contains lots of information (or misinformation) about its circumstances—e.g., *that* it is nearer point A than point B, *that* it is within the boundary of region C, *that* it is between F and G, *that* it is fast approaching agent D, who is on the same path but moving slower, etc. (Notice that I have captured this limited selection of information in a series of "that"-clauses expressed in English.) Some of this information will be utilizable by the robot, we may suppose, and some not. Whatever it can use, it believes (I would say); whatever it cannot use, it does not believe, since although the information is *in* the system, it is not *for* the system: it cannot be harnessed by the system to modulate behavior in ways that are appropriate to the system. Perhaps the fact *that* J, K, and L all lie on a straight line is a fact that we can see from looking at the robot's map, but that the robot would be unable to extract from its map using all the map-reading apparatus at its disposal.

There is a temptation here to think of this map reading or extraction as a process having the map as its "input" and some sentence expressing one or more of these propositions or "that"-clauses as its "output." But no such sentence formation is required (though it may be possible in a talking robot). The information extraction might just as well consist of the generation of locomotory control signals sufficient for taking some

rs
ne
h
ir
2-
"
e
e
o
e
o
r
r
e
e
e
-
g
t
f
a
-
s
l
-
s
-
e

action appropriate to the state of affairs alluded to by the "that"-clause (appropriate, that is, given some assumptions about the agent's current "desires"). That locomotory recipe might not be executed; it might be evaluated and discarded in favor of some option deemed better under the circumstances. But since its generation as a candidate is dependent on the map's containing the information that-*p*, we can attribute the belief that-*p* to the system. All this is trivial if you think about the beliefs a chess-playing computer has about the location and value of the pieces on the chess board and the various ways it might utilize that information in generating and evaluating move candidates. Belief talk can do an acceptable job of describing the information storage and information revision contained in a map system.

Are map systems as versatile as "propositional" systems? Under what conditions does each flourish and fail? Are there other data structures or formats that are even better for various tasks? These are good empirical questions, but if we are going to raise them without confusing ourselves, we will need a way of speaking—a level of discourse—that can neutrally describe what is in common between different robot implementations of the same cognitive competence. I propose the intentional stance (and hence belief talk) as that level. Going along with that proposal means abjuring the inferences that depend on treating belief talk as implying a language of thought.

Alternatively, one could reserve belief talk for these more particular hypotheses and insist on some other idiom for describing what information-processing systems have in common whether or not they utilize beliefs (now understood as sentences in the head). I am not undivorcibly wed to the former way of speaking, though I have made out the case for its naturalness. The main thing is not to let misinterpretations cloud the already difficult arena of theoretical controversy.

There are important and interesting reasons, for example, for attempting to draw distinctions between different ways in which information may be utilized by a system (or organism). Consider the information that is "interwoven" into connectionist nets (as in Janet Halperin's example). As Clark and Karmiloff-Smith (1994) say, "It is knowledge *in* the system, but it is not yet knowledge *to* the system." What must be added, they ask, (or what must be different) for information to be knowledge *to* the system? (See also Dennett 1994.) This is one of the good questions we are on the brink of answering, and there is no reason why we cannot get clear about preferred nomenclature at the outset. Then we shall have some hope of going on to consider the empirical issues without talking past each other. That would be progress.

Do animals have beliefs, then? It all depends on how you understand the term "belief." I have defended a maximally permissive understanding of the term, having essentially no specific implications about the format or structure of the information structures in the animals' brains, but simply

presupposing that whatever the structure is, it is sufficient to permit the sort of intelligent choice of behavior that is well predicted from the intentional stance. So yes, animals have beliefs. Even amoebas—like thermostats—have beliefs. Now we can ask the next question: what structural and processing differences make different animals capable of having more sophisticated beliefs? We find that there are many, many differences, almost all of them theoretically interesting, but none of them, in my opinion, marking a well-motivated chasm between the mere mindless behavers and the genuine rational agents.

REFERENCES

- Clark, A., and Karmiloff-Smith, A., 1994. "The Cognizer's Innards." *Mind and Language*, 8: 487–519.
- Davidson, D., 1975. "Thought and Talk." In *Mind and Language: Wolfson College Lectures*, 1974. Oxford: Oxford Univ. Press.
- Dennett, D., 1969. *Content and Consciousness*. London: Routledge & Kegan Paul.
- Dennett, D., 1971. "Intentional Systems." *J. Phil.*, 68: 87–106.
- Dennett, D., 1983. "Intentional Systems in Cognitive Ethology: The 'Panglossian Paradigm' Defended." *Behavioral and Brain Sciences*, 6: 343–90.
- Dennett, D., 1987. *The Intentional Stance*. Cambridge, MA: The MIT Press/A Bradford Book.
- Dennett, D., 1994. "Labeling and Learning." *Mind and Language*, 8: 540–48. "Learning and Labeling" (comments on Clark and Karmiloff-Smith, *Mind and Language*).
- Fodor, J., 1990. *A Theory of Content*. Cambridge, MA: The MIT Press/A Bradford Book.
- Gregory, R. L., 1987. *Oxford Companion to the Mind*. Oxford: Oxford Univ. Press.
- McCarthy, J., 1979. "Ascribing Mental Qualities to Machines." In M. Ringle, ed. *Philosophical Perspectives in Artificial Intelligence*. Atlantic Highlands, NJ: Humanities Press.
- Nagel, T., 1974. "What Is It Like to Be a Bat?" *Philosophical Review*, 83: 435–50.
- Nagel, T., 1986. *The View from Nowhere*. Oxford: Oxford Univ. Press.
- Premack, D., 1988. "Intentionality: How to Tell Mae West from a Crocodile." *Behavioral and Brain Sciences*, 11: 522–23.
- Searle, J., 1992. *The Rediscovery of the Mind*. Cambridge, MA: The MIT Press.