How to read minds in behavior
A suggestion from a philosopher

Jonathan Bennett

[From A. Whiten (ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (Oxford: Basil Blackwell, 1991), pp. 97–108.]

1. Introduction

Underlying empirical questions about how animals behave, and why, is this: By what formula should we go from premises about behaviour to conclusions about thoughts? In discussing this I shall focus on two thoughtful states—belief and desire.

Neither of those can help to explain behaviour except when combined with the other. Behaviour shows what the animal wants only if we know what it thinks, and conversely. The guiding idea is triangular: cognitive explanations of behaviour are possible only because the animal *does* what it *thinks* will produce what it *wants*. (Henry Wellman's paper in this volume reports that very young children interpret the behaviour of others through a psychology using something like the concept of desire but not the concept of belief. Wellman suggests that the child's conative concept is that of a 'simple want', where 'Sam wants the apple' relates Sam to the apple without requiring him to have any kind of mental representation of the apple. I would rather say this: Whereas a five-year-old predicts what x will do on the basis of what it think x wants and *what it (the child) thinks x believes about the world*, a two-year-old does it on the basis of what it thinks x wants and what it (the child) believes about the world. The very young child is using a belief-desire psychology, but a restricted one in which only the attributer's beliefs are attributed to the subject. This account may even be substantively equivalent to Wellman's own. Either way, there is no conflict between his work and mine.)

So we need to build our account of an animal's beliefs and desires by tackling both at once. That might seem to expose our theory to the risk of vacuity, leaving too unconstrained a choice about what thoughts and wants to attribute to the animal because whatever we say under one heading can be made to fit the behaviour by an adjustment under the other. The animal uttered that piercing scream because it wanted to eat the eagle and thought that the scream would make the eagle fall dead. That is absurd, of course; but we should be able to reject it for some more disciplined reason than that—we need a principled, theoretical protection against uncontrollably free trade-offs between the attributions of beliefs and the attributions of desires.

We are somewhat protected because we can connect attributions of beliefs to facts about the animal's environment. The tie must be mediated by theory about what the animal is sensitive to, but there are ways of checking on that. We are further protected because we can safely attribute to animals desires that are fairly constant across time, except for changes linked to knowable changes in the animal's condition—it wanted food an hour ago but since then it has gorged, and so on. That constancy lets us check attributions of desires at one time against later attributions; and that constrains attributions further, helping to stop the slide into absurdity. If the constancy were not there, and an animal's basic desires changed rapidly with no external pointers to the changes, its behaviour would be unpredictable and therefore unexplainable. For evolutionary reasons, however, there are no such animals.

I shall now take for granted that we have principled ways of avoiding interpretations of animal behaviour that are absurd and obviously not worth entertaining.

That, however, leaves plenty of choices needing to be made, and plenty of disagreements about them. I want to clarify what is at stake in those disagreements and help to resolve them. When front line workers on animal cognition disagree about what states of mind are revealed by what behaviour, they often seem not to agree about what evidence *would* settle the disputes. Everyday working and arguing standards seem to be insecure and idiosyncratic, and it's in that situation that I shall offer some possibly helpful ideas.

2. The economy rule

One popular methodological idea is the view that we should always explain behaviour as economically as possible: don't attribute cognitive states to an animal whose behaviour you can explain without invoking them, and in your cognitive attributions don't go 'higher' on the scale than is needed to explain the behaviour. This 'economy rule' condemns saying that when the chimpanzee Sherman made the sign for a rake he wanted Austin to *think that he (Sherman) wanted the rake*, because the behaviour could as well be covered by supposing that Sherman merely wanted Austin to *bring the rake*. That condemnation seems right, and probably a lot what goes wrong in psychology and cognitive ethology comes from the kind of interpretative overreaching that the economy rule forbids.

However, the rule could not do all our work for us. If we have competing cognitive explanations that do not differ in complexity, sophistication, or whatever it is that feeds into the notion of 'higher', the rule is silent. There, at least, we need more theory.

Even where the economy rule does have something to say, we should ask for its credentials. Why should we always accept the 'lower' or more economical of two unrefuted explanations? Is it because we should always assume things to be homogeneous or unstructured unless we have positive evidence to the contrary? Why should we believe that?

Having accepted the economy rule for years, I now think that we have mistaken its status, and that no deep truth underlies it. I shall justify this in the next two sections.

3. The economy rule as advice

If the rival explanations are not empirically equivalent—if they predict different behaviour—then we should look for or try to elicit further behaviour that fits only one of them. Suppose we are trying to decide whether, when the animal screams like that, this is because it wants others to climb trees or because it wants them to think there is a leopard nearby. Then we should simply try to find out which of these is right. Those hypotheses differ in what they imply

for the animal's behaviour, I shall argue, and that behaviour should be the final arbiter. The economy rule does no real work here. Still, it might function as good advice, telling us to expect that the behavioural data will eventually favour the 'lower' rather than the 'higher' hypothesis, and perhaps advising us to adopt the former as our provisional opinion until the facts are in. This may be generally good advice, but only because on our planet most mentality happens to be fairly low-level. There could be planets where most vaguely goal-seeking behaviour really did involve cognition, and high-level cognition at that; on such planets the advice issued by the economy rule would be bad.

Because I am interested in the foundations of the activity of attributing cognitive states on the strength of behaviour, I exclude linguistic behaviour. Much cognition is expressed without language, and we need to understand how. Also, even where language is present, we can recognize it as language—can know that the speaker means something—only because we can, independently of language, discover things about what he thinks and wants. (Or so I argue in Bennett, 1976.) But although I exclude language, my 'animals' include very young humans; and where *they* are concerned the economy rule's advice—'Expect the "lower" hypothesis to be right'—may be bad.

Andrew Whiten has remarked that the same might be true for chimpanzees. I agree. Some observed chimpanzee behaviour certainly 'feels' like an expression of fairly high-level cognition (see for example the best anecdotes in Whiten and Byrne, 1988), and even where one can produce a 'lower', deflating explanation of the data (as in Bennett, 1988) we may reasonably suspect that in some cases a 'higher' explanation is right. Eventually, however, suspicion should give way to firm evidence.

That, incidentally, will often require not merely hands-off observation of animals' natural behaviour but also conduct that is elicited from the animals by experimentally rigging their environments and their experience (cf. the papers by Verena Dasser and David Premack, and by Dorothy Cheney and Robert Seyfarth, in this volume). Experiments involve certain theoretical risks, which are the price for great practical advantages. Hands-off work is the best if it can be done; but I conjecture that definitive answers to our questions will always require experiments. For good remarks on this, see Dennett, forthcoming.

4. Empirically equivalent rivals

There can be rivalry between hypotheses which, though one goes 'higher' than the other, are empirically equivalent. The 'higher' one must include something explaining why the extra psychological capacity is not used. The lower one might be:

L: The animal has the concepts of one, two and three, and the concept of equal-numberedness, but not the number four,

and its rival:

H: The animal has the concepts one, two, three, four, and equal-numberedness, but it cannot use its concept of four except in doing number comparisons between quartets and other groups.

How could we choose between these? Well, H credits the animal with two more items than L does—namely an extra concept, and a blockage to its exercise—so it makes the animal more complex than L does. Whether we should accept H depends on whether we can justify the extra complexity.

What would justify it? Well, in developing a theory of the animal's internal cognitive dynamics—about how some changes in its beliefs lead to others—we might find that our smoothest explanation for its grasp of one, two and

three implies that it also has the concept of four; and we might have evidence for its having a natural class of cognitive obstacles that would include an inability to employ four except in that one way. In that (admittedly fanciful) case, we should prefer H; but without something like that L should be preferred, not because it is lower but because it is less complex and greater complexity is not justified. This coincides with what the economy rule says, but it comes not from that rule but from perfectly general considerations about simplicity and complexity.

5. A first stab at answering my question

Faced with rival hypotheses that have different empirical consequences, I said, we should get evidence that knocks out one of them. That is easier said than done. Even harder than devising and conducting the tests is figuring out what would count as evidence for or against an hypothesis. That was my initial question: What behaviour indicates what states of mind? So far, all I have done is to take the economy rule down from the throne, while not banishing it from court. Let us start again.

Any explanation of animal behaviour is answerable to a *class* of behavioural episodes. If we have only one episode to go on, we can interpret it only by guessing what *would* happen on other relevant occasions. I shall assume henceforth that we are always trying to explain a longish sequence of behaviours, trying to bring them all under a single explanation. Suppose we have observed a class of behaviours of which something of this form this is true: 'Whenever the animal receives a stimulus of sensory kind S, it engages in behaviour of motor kind M.' For example: Whenever its visual field presents a clear sky with a black patch near the middle of it, and occupying at least 1% of the field, the animal utters a specific kind of noise.

Here are two rival explanations for this behaviour. **(i)** The animal has an innate or acquired stimulus-response disposition; it is hard- or soft-wired to make that noise upon receipt of that visual stimulus. On each occasion in the class it received such a stimulus and accordingly made the noise. **(ii)** The animal has the safety of its group as a goal. On each occasion in the class it thought it saw a predator and called to warn others of danger.

To test **(i)** we should vary the circumstances while still presenting that kind of stimulus, and see whether the animal still gives that call. To the extent that it does, the hypothesis is confirmed. Of course, the call might be triggered by another kind (or other kinds) of stimulus as well. Suppose we discover that the animal does also make such a cry whenever it gets a stimulus of some third kind, then a fourth, a fifth,. . . and on into dozens of different kinds of sensory intake, each leading to the same kind of behaviour. If this happens, we are under increasing pressure to find some unifying account of all this behaviour, some one explanation to replace the multitude of separate stimulus-response ones that we have accumulated.

(a) There might be no way of doing this. **(b)** Or we might find that there is after all a single sensory kind of stimulus on all the occasions—a subtle smell or a high-pitched sound—enabling us to cover all the cries by a single stimulus-response generalization, after all. **(c)** Or we might find that we could bring all the episodes under a single generalization but not a stimulus-response one. Even if no one *sensory* kind of stimulus is shared by all the episodes—no configuration of colour, shape, smell, etc.—they may have something in common that lets us generalize across them, namely the fact that each of them *provides evidence to the animal that there is a predator nearby*. If they share that, and there is no more economical way of bringing them under a single

generalization, that gives us evidence that the episodes are united in that way for the animal itself. That is tantamount to saying that in each episode the animal thinks there is a predator nearby.

What entitles us to bring the proposition *There is a predator nearby* into our description of the animal (through the statement that that's what it believes) is our having a class of behavioural episodes that can be united with help from the proposition *There is a predator nearby* and cannot be united in any simpler way. Our best unitary account says that in each environment where it calls *the animal has evidence that there is a predator nearby*. (What about 'There is a predator nearby'? That fact could not immediately help to explain the animal's behaviour. No fact about the environment could explain its behaviour except by being somehow registered upon or represented within the animal's mind.) The fact that we can unify the occasions with help from an embedded 'that P', and in no other way, justifies us in using an embedded 'that P' in explaining the behaviour.

I don't know a long history for the 'unification' idea proposed here, though it may have one. I propounded it in Bennett, 1964 (section 2) and Bennett, 1976; it is put to good use in Whiten and Byrne, 1988, where acknowledgment is also made to Dawkins, 1976.

The proposal is not merely about when we may explain behaviour by attributing beliefs, but also about what beliefs we may attribute. We get at belief content through what is perceived as common to all the environments in which the behaviour occurs. I shall return to this, the central theme in my paper, shortly.

First, a small correction is needed. The basic belief-desire-behaviour story must focus on beliefs about means to ends, that is, about what movements on the animal's part will bring about what it wants. We can attribute beliefs of other sorts—e.g. that there is a predator nearby—only through attributing beliefs about means to ends, which alone are immediately tied both to wants and to behaviour. When, therefore, we hypothesize that the animal calls because it thinks *there is a predator nearby*, that should be based on the hypothesis that it calls because it wants its companions to be safe and thinks that *that cry is a means to their being safe* because it thinks there is a predator nearby.

6. Thoughts about thoughts: preliminary tidying

What would count as behavioural evidence for us that our animal has a thought about some other animal's mind? This thought could be either a belief or a desire, and what it is about could be either a belief or a desire.

Or an animal might have a thought about another animal's perceptual state. Our subject animal might behave in a certain way because of what it thinks about what another animal might hear or smell or otherwise take in. Supposed evidence for that kind of thought is often misleading. Usually, behaviour that is supposed to manifest the thought 'This will stop x from seeing y' or 'This will stop x from smelling y' could just as well be manifesting the thought 'This will put a physical object between x and y' or 'This will put y upwind from x'. How to get good evidence for thoughts about perceptions or sensory states is an interesting question, but I shan't discuss it here. (The paper by Cheney and Seyfarth in this volume says interesting things about it.) My topic is the more ambitious attribution of beliefs or desires whose topic is other beliefs or desires.

Such an attribution might fit into our explanatory schema in various ways. Here is one: We have evidence that our animal wants to achieve goal G and thinks that doing A will bring this about; and we don't see how it could arrive at that belief except through attributing a certain mental state to

some other animal. For example, we don't see how it could think its cry will make its companions safe except through thinking that the predator *wants* to eat them.

It will be hard to make that stick, however. Almost certainly, we shall be able to explain the basic attribution through the animal's thinking merely that the predator *will* eat the others if it catches them. That is, a supposed belief about a want should give way to a belief about a simple behavioural disposition if the latter covers the data as well.

I shan't discuss what would entitle us to attribute a belief about a predator's desire. I choose for detailed discussion the attribution of a desire to produce a belief. My treatment of that will indicate how in outline I would deal with the other belief-desire combinations.

7. Desires to produce thoughts: a dilemma

What sort of evidence could entitle us to hypothesize that our animal behaves as it does because it *wants to produce a thought* in its companions, e.g. because it wants to get them to think there is a predator nearby?

It is highly improbable, in the nonhuman world, that our animal should want its companions to have the belief that P just for itself, as an intrinsic good. Let us focus on the less wild possibility that our animal wants its companions to believe that P because of how that belief will affect their behaviour. For example, it calls so as to get them to think there is a predator nearby, which it wants as a means to their behaving thus and so.

To be entitled to say that, we must rule out everything like this: The animal calls so as to get its companions to crawl under a bush. If they do behave thus whenever it calls, our animal may see its call as a trigger to produce the crawling, with no thought of what its companions will think. If it sometimes calls when its companions are already under bushes, that doesn't help, for it might always call so that it companions will *be* under a bush—going there or remaining there.

Objection: 'If our animal thinks that the cry will elicit the hiding behaviour, it must have some belief about *why* it will do so; and the most likely candidate for this is a belief that the cry will cause the others to believe that there is a predator nearby.' This presupposes that a means-to-end belief must be accompanied by a belief about why that means leads to that end; which is absurd. Across the centuries most human means-to-end beliefs have been merely empirical—accepted without any grasp of why they are true, simply because they are confirmed by past experience. If we can do that, why not other animals?

If our animal is to be credited with wanting to produce not merely behaviour but a belief in its companions, the evidence must be enriched—but how? We have to suppose that our animal wants to give the others a belief as a means to their using it in their behaviour, but we don't want

> evidence that: our animal calls as a means to producing a belief which it wants as a means to producing behaviour

to collapse into

> evidence that: our animal calls as a means to producing behaviour,

with the intended belief dropping out, not attributed because there is no work it needs to do.

8. A way of escape

I know only one solution to this dilemma. Suppose that in the series of episodes when our animal calls, its companions act variously, depending on their states and situations: if they are F, they run; if they are G, they search; if they are H, they freeze; if they are J, they climb; if they are K, they dig,. . . and

so on; and whatever each animal does is appropriate to the information that there is a predator nearby. Can we still interpret our animal's purpose in calling as merely to elicit behaviour? Here is how such an interpretation would go:

> The animal's past experience has shown it that when it calls like that its companions run if they are F or search if they are G or freeze if they are H or climb if they are J or dig if they are K or. . . ., and on this occasion it wants them to run if they are F or search if they are G or freeze if they are H or climb if they are J or dig if they are K or. . . ., and so it calls.

This is now crediting our animal with a thought of implausible complexity. We can simplify the story and make it more credible by supposing that our animal unites the complex thought

> run if they are F or search if they are G or freeze if they are H or climb if they are J or dig if they are K or. . . ,

into the unitary thought

> behave appropriately to the fact that there is a predator nearby.

That brings their behaviour under a description—call it D—that has nested within it the complete proposition that *there is a predator nearby*. If D is our best way of unifying all the behaviours of the group, that is evidence that on those occasions the animals believe that there is a predator nearby. And if D occurs in our simplest statement of what our subject knows about its companions and of what it wants, that is evidence that when it calls it does so because it wants them to believe there is a predator nearby.

9. A further difficulty resolved

This is still not right, however. In the story as I have told it, the calling animal's success on each relevant occasion consists in this: Its companions don't get eaten. Suddenly we slump back into a simple story that is purely physicalistic once more, and not psychological. What it knows from past experience is that if it gives that cry when there is a predator nearby, its companions don't get eaten; it doesn't want them to get eaten now when there is a predator nearby; so it calls again. That is not horrendously complex, and it does not credit our animal with a thought about a thought.

So we are back at square one! But if my general strategy has been right, we can see what would in principle deal with this latest trouble and—at last—have a chance of keeping us out of trouble. What is needed is that the animals have a variety of uses for the information that a predator is nearby. There is no hope of that, of course, so let us switch from predators. Suppose there is some other kind of object—call it a Quark—which our animals can use in different ways, depending on their condition and circumstances: they can eat it, shelter under it, use it to crack open coconuts,. . . and so on. If the range of appropriate responses to the information that there is a Quark nearby is sufficiently various, and if it doesn't all come together again in some one upshot of all these different activities (like escaping the predator), then we can say that the calling animal calls so as to get the others to think there is a Quark nearby. Without such variety, I can find no justification for attributing to our animal any desire to produce a belief.

I have brought us to a point that may seem to lie beyond anything that is true of actual nonhuman animals. If so, then I am committed to saying that we shan't ever get good evidence that any nonhuman animal wants to produce a belief. Maybe we could still have evidence that animals sometimes want to produce desires, or have beliefs about beliefs or desires; though I suspect that those would be no easier than the other. I don't know how pessimistic to

be about this: it is early days yet, some observations of chimpanzee behaviour (at least) are suggestive, and there is much more experimental work to be done.

10. A final unsolved problem

I have described a procedure for deciding what mental content to attribute to an animal. I have offered contentions of the form: If an animal's behaviour is thus-and-so, such-and-such thoughts can be attributed to it; and secondly I have suggested that the attributions will be unjustified if the animal's behaviour does not conform to the patterns that I have described. Never mind the second bit; let us focus on the first, which says that my procedure is correct, so that adherence to it will reliably lead to true attributions of thoughts to animals.

If my procedure is in that sense correct, *why is it correct?* What is the logical status of truths of the form 'If the animal's behaviour exhibits this and that input-output pattern then it has such and such mental content'? Two broad kinds of answer can be distinguished.

(a) When we attribute a belief or desire to an animal, we are saying something about its inner state—something that goes beyond any facts about how it behaves—and the facts about its behaviour are merely reliable pointers to those inner states.

(b) When we attribute a belief or desire to an animal, *all* we are doing is to say something complex about its patterns of behaviour. The behaviour is not evidence of the animal's having inner states of belief and desire; rather, to behave like that *is* to have beliefs and desires.

There is much controversy between the adherents of the two positions. In earlier decades, the friends of **(a)** would characterize beliefs and desires as 'mental' in some way that puts them outside the physical world. This Cartesian view has fallen into deserved disfavour, but **(a)** still has its friends, who say that what makes it the case that an animal thinks that P or wants G is some fact about its brain-state and not about how it behaves.

This has its attractions, and just twice in this chapter I have allowed myself turns of phrase that align me with it.

In section 5 I wrote that if a class of behavioural episodes share some feature F, 'and there is no more economical way of bringing them under a single generalization, that gives us evidence that the episodes are united [by F] for the animal itself'. That implies that when the animal has (whether in a belief or a desire) a thought that is applicable to a variety of situations, it really does *have something* that *enables* it to treat all those situations in one same way. This implies a kind of inner realism about mental content. If I wanted not to commit myself to that, and to remain free to give answer **(b)** to the status question, I ought to have written not 'that gives us evidence that the episodes are united in that way for the animal itself' but rather 'that entitles us to avail ourselves of that unity in what we say about what the animal thinks'. For a **(b)** theorist, the concepts of belief and desire are conveniences, aids in the management of certain complex facts about animal behaviour, but they don't have to correspond to items in the animal which enable it to manage its complex data.

In section 8 I wrote that if in a certain case we credit an animal with a thought about behaviour but not one about thoughts, we must credit it with a thought of implausible complexity. That was in the spirit of Premack's statement, quoted in the Introduction to this volume, that 'The ape could only be a mentalist. . . , he is not intelligent enough to be a behaviorist'. My remark and Premack's both imply that when we credit an animal with a simple thought about

a thought rather with a complex thought about behaviour, our attribution doesn't merely apply a conceptualization that serves our theoretical purposes but credits the animal with having a simplifying *something* inside it, a something that makes its behavioural data more manageable to it.

Well, I do sometimes find it natural to write like that. But at other times I am not so sure. Suppose we discovered for sure what enabled the animal to engage in the complex behavioural pattern on the basis of which we have attributed the belief that P. Suppose, specifically, that we found that this pattern of behaviour was possible simply because the animal's brain contains thousands of different though interrelated mechanisms, each dealing with a different input-output pair, and that no one item in the animal was in any way responsible for the belief-manifesting pattern. If we knew all that, would we still be willing to attribute to the animal the belief that P? Sometimes I am strongly inclined to answer 'Why not?', which aligns me with answer (b) to the status question.

This issue is discussed in the Introduction to this volume. The present chapter raises it, but has made no attempt to answer it.

* * * * * * *

Bennett, Jonathan 1964 *Rationality*: London: Routledge and Kegan Paul. Reissued in 1988: Indianapolis: Hackett.

Bennett, Jonathan 1976: *Linguistic Behaviour*. Cambridge University Press. Reissued in 1989: Indianapolis: Hackett.

Bennett, Jonathan 1988: Thoughts about thoughts. *Behavioral and Brain Sciences*, 11, 246f.

Cheney, Dorothy L. and Robert M. Seyfarth 1991. Reading minds or reading behaviour? Tests for a theory of mind in monkeys. A. Whiten (ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (Oxford: Basil Blackwell, 1991).

Dawkins, Richard 1976: Hierarchical organisation: a candidate principle for ethology. P. P. G. Bateson and R. A. Hinde (eds), *Growing Points in Ethology*. Cambridge University Press. pp. 7–54.

Dasser, Verena and David Premack 1991: The emergence of metarepresentation in human ontogeny and primate phylogeny. Whiten (ed.) op. cit.

Dennett, Daniel C. forthcoming: Out of the armchair and into the field. *Poetics Today*, Israel.

Wellman, Henry M. 1991. From desires to beliefs: Acquisition of a theory of mind. Whiten (ed.) op. cit.

Whiten, Andrew and Byrne, Richard W. 1988: Tactical deception in primates. *Behavioral and Brain Sciences*, 11, 233–244.