Cognitive ethology: Theory or poetry?

Jonathan Bennett

from *Behavioral and Brain Sciences* 6 (1983), pp. 356–358.
Comments on Daniel Dennett, 'Intentional Systems in Cognitive Ethology: the "Panglossian Paradigm" defended'. *Behavioral and Brain Sciences* 6 (1983), pp. 343–345.

Dennett is perhaps the most interesting, fertile, and challenging philosopher of mind on the contemporary scene, and I count myself among his grateful admirers. But this present paper of his, enjoyable as it is to read, and acceptable as its conclusions are, is likely to do more harm than good. Some will object that the intentional stance is a dead end; but I think, as Dennett does, that it is premature to turn our backs on explanations of animal behavior in terms of desires and beliefs, and I am in favor of continuing with this endeavor; but only if it gets some structure, only if it is guided by some firm underlying theory. That is what the ethologists might get from philosophy, but Dennett has invited them to turn their attention toward philosophy only to give them a mildly upgraded version of the unstructured, opportunistic, rambling kind of thing they are doing already. He encourages them to go on believing that the conceptual foundations of cognitive ethology are rather easy to lay—a few broad strokes of the brush, or slaps of the trowel, and there you are. Really, it is much harder and more laborious than that. I shall sketch the sort of thing that is needed, and point out some things in Dennett's paper that suffer from the lack of any proper foundations.

I take it as uncontroversial that the intentional stance—considered as a program for theorizing about behavior—must be centered on the idea that beliefs are functions from desires to behavior, and that desires are functions from beliefs to behavior. Down in the foundations, then, we need some theory about what behavior must be like to be reasonably interpreted as manifesting beliefs and desires; and these concepts must presumably tail off somehow, being strongly applicable to men and apes, less strongly to monkeys, and so on down to animals that do not have beliefs and desires but can be described in terms of weaker analogues of those notions. What will this 'tailing off' look like? In my attempt to answer this (Bennett 1976) I have taken it that a theory of belief and desire will be nested within a broader theory of *goals*, or of a *teleological explanation of behavior*. The basis for the latter is to be found, I think, in Taylor (1964), which highlights the idea of what I call an 'instrumental property' of an organism, that is, a property of the form: 'x is so situated and constructed that if it soon does A it will become G shortly thereafter'. Let us put this by saying, for short, that the animal is A/G.

Teleological explanations come into play only if we have a system (e.g. an animal) regarding which there is a reliable generalization of the form: 'For any *a*, whenever it is *a*/G it proceeds to do *a* if that is within its physical competence.' If the animal has eating as its G, its goal, it will dependably kill when 'it is killing/eating, climb when it is climbing/eating, and so on. I am suppressing many complications, but one must be faced openly. No actual animal will, for any G, do whatever *will* bring it G. You give me an animal and a value of G for which this is supposed to be true, and I will rig a situation in which the animal will get G if and only if it lies down and then stands up, three times in quick succession (e.g. I will decide to give it G if and only if it behaves in that way). But it won't act like that unless I somehow inform it of the relevant fact about its situation. So a theory of teleology that is to have any chance of fitting actual animals must rest on generalizations not of the form 'If it is *a*/G it will do *a*' but rather 'If it *has the information that it is* a/*G* it will do *a*'.

In my book I coined the term 'registration', speaking of the animal's being *a*/G as a fact that may be 'registered' upon it; and then I argued that belief is a species of registration, the differentia being a matter of degree which I tried to describe. I probably didn't get it right, but that is of no great moment. What marks off the genus 'goal' from the species 'desire' or 'intention', and the genus 'registration' from the species 'belief', is far less important than is the structure of the genus. That is, what matters is to have a good theory of teleologically explicable behavior, with the foundations of a theory of cognition embedded in it. And I offer my attempt at this in Bennett (1976) not as a source of the right answers, perhaps, but as a fair indication of what some of the principal questions are. I contend that something of that general nature—and not less complex than that—is needed as a foundation for the intentional stance, if the latter is to be worth anything as theory, rather than merely expressing a liking for one way of talking, a kind of dim poetry.

The most important thing in any foundational theory will be its answer to the question, What makes it all right to explain an event teleologically, bringing it under a generalization of the broad form of 'If x registers that it is *a*/G it does *a*, for any *a* within its physical competence'? If the event could be explained in that way and no other, that would justify using the teleological explanation. But what if every event can be explained mechanistically, i.e. in terms of its subject's intrinsic properties, with no mention of any property of the form A/G? I answer that it is all right to bring x under a teleological generalization if the latter captures a class of events that is not covered by any *one* generalization of a mechanistic sort. Where there is a contest between one teleological and one mechanistic generalization (or even, perhaps, two or three of the latter), mechanism wins because it is more basic, uses concepts of wider applicability, and so on (see Taylor 1964, p. 29). But if a teleological generalization does work for us—giving us classifications, comparisons, contrasts, patterns of prediction that mechanism does not easily provide, then that justifies us in employing it. This, I submit, is the *Grundgesetz* of the whole theory of teleological explanation, and thus of the intentional stance.

It bears heavily on one of Dennett's theories. He rightly says that any attribution of beliefs and the like to an animal must be able to stand its ground against lower-level 'killjoy' rivals, and he gives some nice examples of attributions withdrawn—or behavior 'demoted'—in the light of further evidence. This can happen not only when a high-level attribution is challenged by a lower-level one ['Does he want me to think he is hungry, or only to give him food?'] but also when a lowest-level intentional attribution is challenged by a rival that docs not involve intentionality at all.

There is a problem about the latter kind of issue, which Dennett describes but does not explain. Suppose we are inclined to think that a certain animal has as a goal escaping from leopards. That is, whenever its being *a*/escapes-from-leopard is registered on it, it does *a* (subject to complications and qualifications which I shall continue to omit). What would a challenge from below. a killjoy rival, look like in such a case? It would consist in the discovery that the class of events that we had brought under a teleological generalization could also be brought under a non-teleological one. For purposes of this particular point I shall simplify the teleological form even further, and take it to be: 'If the animal (registers that it) is in a leopard-threatening situation it does a leopard-avoiding thing.' We might opt for that generalization—or for the teleological one of which it is a simplified caricature—because we could find no principle of unity for that class of events except the one provided by 'leopard-betokening' in the input and 'leopard-avoiding' in the output. But now suppose we discovered that there is a kind of stimulus S and a kind of behavior R such that **(i)** S is definable without help from any concept like that of 'being evidence for' or 'registering' (e.g. S is a kind of smell, definable in purely chemical terms), and **(ii)** R is definable without help from any concept like that of 'tending to' or 'being apt for' (e.g. R is a motor kind of movement, definable in terms of how certain muscles are used), and **(iii)** the class of supposedly leopard-avoiding situations also falls under the generalization that whenever the animal receives an S stimulus it emits an R response. In that case, the generalization 'Whenever it is in (what it registers as being) a leopard-threatening situation it does a leopard-avoiding thing' should be relinquished: The intentional stance has no honest work to do here, because all its work is equally done by something that is preferable to it because lower level. (Whether the S-R pattern is hard-wired or a result of learning is quite irrelevant, so far as I can see.)

Now, Dennett sees intentional explanations of behavior as threatened by stimulus-response rivals, but he does not say why, except to remark that 'the acts that couldn't plausibly be accounted for in terms of prior conditioning or training or habit [are the ones] that speak eloquently of intelligence' and thus of intentionality. If Dennett wants to be really useful to cognitive ethologists and psychologists—giving them what they need rather than what they want—he ought not to be talking in this way about what 'speaks eloquently' of what, nor should he rely on the term 'training', trusting his intended audience to understand how the kind of training that does not require intentionality differs from the kind of learning that does. Rather, he should be helping them to understand what conceptual structures are involved here. That would require him to have much more theory than he has. He would have to descend from the level of sweeping remarks about stances and levels, and talk in detail about how the levels relate to one another.

This lack of theoretical structure goes very deep in Dennett's paper—right down to the level of the question of what intentionality is. Apart from giving its nominal essence by saying that it is the home ground of intentions, beliefs, and the like, Dennett mentions only one thing that can 'mark' the sphere of the intentional, namely that it involves referential opacity. But the converse doesn't always hold: Some opaque contexts are not intentional; and in any case, how is our grasp of intentionality supposed to be *helped* by this mention of opacity? It has nothing to offer to the floundering ethologist or psychologist, and Dennett makes no use of it in the subsequent discussion. He did need to say something of a technical nature about intentionality, but not *that*. What was needed was rather an account of intentionality as the locus

of one kind of function from sensory inputs to behavioral outputs of animals: A description of what those functions are, of how they actually work, would have meshed with things that ethologists and psychologists do, helping them to get somewhere with their problems; whereas what Dennett says about opacity does not turn any of the wheels that badly need turning at present.

The absence from Dennett's paper of any theoretical underlay also makes itself felt in his treatment of what he sees as a problem confronting anyone who wants to base intentionalist conclusions on ethological data. The problem, according to Dennett, is that the best evidence for intentionality comes from what an animal does in unusual circumstances; such evidence will take the form of relatively isolated anecdotes; and trained observers are taught to be wary of anecdotes, and to concentrate on getting hard data, that is, oft-repeated patterns of behavior. So there is a danger that accepted canons of good scientific conduct will act as a sieve, keeping the best evidence for intentionality from getting through onto the pages of the observer's log book. For this difficulty, he offers two solutions: **(i)** We can 'pile anecdote upon anecdote, apparent novelty upon apparent novelty', until it becomes incredible that there is not a real underlying intentional pattern. **(ii)** We can devise experiments, set traps, and so on, trying to provoke 'novel but interpretable behavior', thus '*generating anecdotes* under controlled (and hence scientifically admissible) conditions'.

I object that Dennett has not explained why the problem exists, because he has not said why non-anecdotal evidence cannot support attributions of belief and desire, except for remarking that it does not 'speak eloquently' of intentional states and may be explainable in terms of 'conditioning or training or habit'. I also object that he does not explain why his proposed solutions *are* solutions, or, for that matter, how they are to be executed. He does not say what kind of 'pile' we should heap up in solution **(i)**, and in **(ii)** he leaves it unclear how the poison of anecdote is supposed to be neutralized by the antidote of control.

In fact, his problem arises only if observers are looking for behavior that can be brought under generalizations relating sensory kinds of input to motor kinds of output, for example, saying that when the animal encounters a certain kind of smell it moves certain muscles thus and so, rather than generalizations relating evidential kinds of input to consequential kinds of output, for example, saying that when the animal encounters signs of the proximity of a leopard it does something that is apt to get it out of the leopard's vicinity. Suppose we have an animal that whenever it encounters an S smell makes R movements; and suppose that usually an S smell is evidence of leopards and R movements do provide escapes from leopards. Now, we are wondering whether this behavior, conforming as it does to an S-R pattern, should be explained intentionally, that is, brought under the generalization that when the animal is (or perceives itself as) leopard-threatened it leopard-avoids. To find the answer, we must vary the conditions, bringing it about that the animal sometimes gets evidence of leopards other than S smells, and sometimes needs something other than an R movement to avoid a leopard; and we must observe how it behaves in these situations, either on a first encounter or after a number of trials from which the animal can learn things about evidence for leopards and means of escape from them. If the 'leopard' generalization holds good in cases in which the S-R generalization fails, or in cases in which it is inapplicable, that helps to justify our using the 'leopard' generalization, which is tantamount—given the simplification with which I am now working—to bringing the intentional stance to bear on the behavior in question. Despite what Dennett says, this is not a

move from regularities to anecdotes; rather, it is a move from regularities of one kind to regularities of another. If the work is done right, there may indeed be 'control', but that is not what makes the procedure 'scientifically admissible'. There is no reason in principle why we should not make the enlarged set of observations with our hands behind our backs, not contriving anything but just looking in the right direction. The procedure is scientifically admissible just because it consists in objectively attending to data in the light of a decent hypothesis; and it bears on intentionality because of what the hypothesis is. I think, as Dennett evidently does, that the ethological and psychological literature contains little convincing evidence of non-human intentionality. But that is not because intentionality is inimical to regularity and thus to normal scientific method; rather, it is because the people doing the work don't know what regularities to look for, having no theory of intentionality. I am afraid that Dennett's paper will encourage them to go on being content to have none.

Theoretical foundations are needed not only along the borderline between intentional and non-intentional, but also in adjudicating between a given intentional hypothesis and some lower-level intentional rival to it. Consider, for example. the contrast between 'Tom wants Sam to believe that there is a leopard' and 'Tom wants Sam to run into the trees'. Dennett rightly implies that behavioral evidence can discriminate between these, but his only suggestion about how it can do so is wrong or seriously incomplete. He handles 'Tom wants Sam to run into the trees' in terms of Tom's using a 'trick' to 'induce a certain response in Sam', and compares this with getting someone to jump by shouting 'Boo!' at him. The impression is given that a first-order intention must be an intention to trigger an automatic response; but that is just wrong, for we have a first-order intention whenever an animal intends to bring it about that P, where P does not involve any intentional concepts. Thus, Tom may intend to get Sam to run into the trees, and the mechanism that actually operates in Sam may involve an inference from 'Tom wants me to run to the trees, and usually it pays to do what Tom wants me to do' to the conclusion 'It will be worthwhile to run to the trees'. Tom's intentionality is not prevented from being first order by the fact that what happens in Sam—as distinct from what Tom intends or wants to happen in Sam—is itself intentional.

How, then, can behavior mark the difference between 'wants Sam to believe there is a leopard' and 'wants Sam to run into the trees'? Well, I think that it cannot mark the difference unless there are circumstances in which Tom thinks there is a leopard nearby and in which that fact makes it appropriate (relative to Tom's value system) for Sam to do something other than running to the trees. If there is a kind of behavior that Tom engages in whenever he thinks there is a leopard nearby, and if in each instance he behaves with the intention of getting Sam to do A, or do B, or do C, through a long list of kinds of behavior that have nothing in common except their appropriateness to there being a leopard nearby, then, and only then, are we entitled to say that what Tom wants is something describable with the aid of 'There is a leopard in the vicinity'. (I am here applying some thoughts I first developed in Bennett 1964, pp. 19-21.) It may, however, only be 'Tom wants Sam to do something appropriate to the fact that there is a leopard in the vicinity'. To be entitled to say that Tom wants Sam to believe that there is a leopard, we shall need further evidence; and it won't be easy to find. I suspect, indeed, that if we are ever to be entitled to interpret non-human animals in terms of anything higher than first-order intentionality, that will have to be because for non-human animals we adopt specially

weakened intentional concepts. But perhaps not. In any case, whatever concepts are being used, they had better be understood: they had better exist as theoretical items, not as mere predilections for using words in certain ways. Otherwise the entire project will continue to wander in the wilderness.

I wonder what Dennett's picture is of the project as it has been pursued up to now. In a footnote he refers to two of the Yerkes chimpanzees, saying that their 'apparently communicative behavior. . . cries out for analysis and experimentation via the Sherlock Holmes method', that is, through the accumulation of controlled and contrived anecdotes. He does not mention the fact that the Yerkes psychologists think that they *have* analyzed the communicative behavior of their chimpanzees (Savage-Rumbaugh et al. 1978). He must think it is possible to do better than they have, and I agree with that. But what kind of improvement in the analysis does Dennett have in mind? He gives the impression of thinking that adjudicating between rival interpretations is always to be handled in terms of informal, intuitive intelligence as brought to bear on the particular case, and that it shouldn't be very hard to get agreement by such means. If that is his view, then presumably he will think that what philosophy has to offer is just some rough guidance on how to be 'careful' in thinking about cases, some help in getting 'the knack'. I submit that that is far too undemanding a picture of what is needed in adjudicating between rivals. And even if it were not, a proper underlying theory would still be needed to help students of animal behavior to know how to construct rivals to a given hypothesis and how to look for positive evidence that there aren't any rivals. Both sorts of help are desperately needed, judging by the literature to date.

Dennett's handling of the intentional stance—typified by his willingness to describe data in terms of what 'speaks eloquently' and what 'delights' or 'depresses', rather than of what does or does not satisfy explicitly stated criteria—is puzzling. For he declares an interest in developing a 'suitably rigorous abstract language in which to describe cognitive competences', and says: 'We are interested in asking what gains in perspicuity, in predictive power, in generalization, might accrue if we adopt a higher-level hypothesis that takes a risky step into intentional characterization.' Despite a puzzling later remark about the stance as not a theory 'in one traditional sense', he clearly does regard it as enough of a theory to make my criticisms prima facie relevant. I can only suppose that his silence about all the theory's details arises from his thinking that the details are rather obvious and easy. Well, they are not; and much more work must be done on them if the intentional stance is to get anywhere. In implying the contrary, Dennett has misestimated the confusion and conceptual shallowness that reign throughout the relevant literature to date. And he has also misestimated his own needs in this very paper, as I have tried to show in pointing to some (not all) of the things in the paper that would have gone better if some explicit theory had been at work.

References

Bennett 1964: Jonathan Bennett, *Rationality: an Essay Towards an Analysis*, Routledge.

Bennett 1976: Jonathan Bennett, *Linguistic Behaviour*, Cambridge U.P.

Savage-Rumbaugh et al. 1978: Sue Savage-Rumbaugh, Duane M. Rumbaugh, and Sally Boysen, 'Linguistically mediated tool use and exchange by chimpanzees (Pan troglodytes)', *Behavioral and Brain Sciences* 1.

Taylor (1978): Charles Taylor, *The Explanation of Behaviour*, Routledge.