Analysis without Noise

Jonathan Bennett

‘Analysis Without Noise’, in Radu Bogdan (ed), *Mind and Common Sense* (Cambridge University Press, 1991), pp. 15–36.

1. Folk psychology and conceptual analysis

This paper will present some conceptual analysis, trying to command a clearer picture of how our mentalistic concepts work. I mean our untutored workaday concepts, the ones that we employ in folk psychology. Many students of the mind these days are poised to jettison folk psychology as soon as they can, in favor of something better informed, more comprehensive, more closely in touch with the central nervous system, or the like. But given that folk psychology, whatever its defects, is not on a par with alchemy or astrology, we oughtn’t to drop it until we understand more than we do about what kind of theory it is, what work it does, and how. That’s one reason why even the iconoclasts should be interested in that part of the philosophy of mind that consists in old-fashioned conceptual analysis. A second reason also weighs with me: folk psychology is a wonderful intellectual construct, an amazing tool for enabling us to get on with one another, to manipulate and predict one another, and to evade and foil such manipulations and predictions; it is an inherently worthy object of study.

The network of concepts that we use in folk psychology pretty well exhausts its content, considered as a compendium of general propositions: it is hard to find much universally received general doctrine about the human mind that doesn’t qualify as highly analytic. In treating analyticity as a matter of degree, I side with Quine’s view that there is a continuum between sentences that are true by virtue of their meanings and ones that are universally accepted as expressing well entrenched truths about how the world is.[1]

That there is so little in folk psychology that counts as clearly contingent is easy to explain. Firstly, folk psychology is an old theory, so that its principal doctrines are deeply entrenched in culture, language and literature. Secondly, it is a largely unchanging theory: within historical time, it seems, there has been virtually no come and go in its content. Because of this doctrinal stasis, nothing has happened to force us to distinguish the more from the less deeply entrenched parts of the theory, i.e. to make discriminating decisions about *how* to accommodate facts about the mind that are seen to be recalcitrant vis-à-vis the totality of accepted doctrine. These two facts combine to produce a situation in which virtually the whole of folk psychology is considerably analytic.[2]

That makes it hard to see that folk psychology is a theory at all. A theory is presumably something that could be found to be false, and when we try to envisage discovering the falsehood of any part of folk psychology we run into a kind of absurdity: the challenged item turns out to be so deeply entrenched as to count as analytic, which makes us say that *it* couldn't be false, though we might come to use those words to express something that is false. Of course this doesn't make folk psychology invulnerable; it merely replaces the question 'Might any of its theses be found to be false?' by 'Might all or part of the theory be found to be inapplicable?'

2. The noise problem

One way of doing conceptual analysis is to present oneself with stories using the terminology under investigation and ask oneself whether they are 'intuitively' acceptable. If they are not, that fact helps us to understand our concepts by setting some limits to what they will tolerate. Consider, for example, the following two-part story:

(1) System S takes in signals from its environment and responds with physical behavior, all of this happening according to a *mind-indicating pattern*, meaning a set of (input→output) conditionals such that a system's richly conforming to that set would give it as much entitlement to count as a perceiver/thinker/wanter as input-output relations could possibly give it.

(2) S's behavior falls into a mind-indicating pattern only because it is being manipulated by someone who acts on the following plan: *I shall bring it about that S's behavior never conflicts with conditionals. . .* , where the blank is filled by a list of all the conditionals that define the system's mind-indicating pattern of behavior. The manipulator thinks of the conditionals just as a list, and does not know that they constitute a mind-indicating pattern.

Is S a perceiver/thinker/wanter? Searle thinks not, and labors to get readers to agree with him.[3] These attempts of his are not arguments but appeals to intuition. The whole performance fits the schema: Tell a story and then ask whether our concepts tolerate it.

This is apt to be a risky procedure because of the *noise problem.* Why do we gag at the suggestion that S is a thinker? Is it because we have a concept of *thinker* that excludes S as described, or is it rather that this bizarre story is something for which we are not forewarned or, therefore, forearmed? Are we rejecting it in a controlled and disciplined way, or are we merely being knocked off our conceptual pins by it? Testing one's conceptual intuitions against this story about S is like testing a coin by ringing it on the counter of a busy boiler factory: there is far too much noise for a trustworthy signal to come through. We need to filter out the noise—the bizarreness reactions that are not informative about conceptual structure—if we are to know whether we have here any evidence about the limits to our concepts, i.e. about the structures through which we actually think about thought.

3. The thin end of what wedge?

The procedure is also dubiously useful. Suppose that our mentalistic concepts really do entail that S is not a thinker because it does not satisfy some *inner-route constraint*, some conceptual requirement on how thinkers, properly so-called, must get from inputs to outputs. What makes that information worth having?

Well, it refutes those philosophers who hold that our mentalistic concepts are purely externalistic in their demands, so that the right kind of input/output pattern is enough to qualify a system as a thinker and wanter. But nothing is so boring and trivial as the information that someone has believed a falsehood. I want to know what adult reason there is for caring whether our concepts include a barrier to counting S as a thinker and wanter. What reason is there that has to do with the advancement of understanding? There may be none.

For one thing, the information that our concepts do include such a barrier is enormously thin. Look at it:

> Our concepts won't classify as a thinker something that fits the right input/output patterns only because it is caused to do so a thinking manipulator who is guided by a list of conditionals and is not aware of the patterns as such.

Because the story about S is a rather strong, rich one, its negation is relatively thin and empty. The information that our conceptual scheme requires its negation is, therefore, correspondingly thin and empty. It's not a null result, but it comes close.

Of course, this tiny discovery might be the thin end of a wedge; perhaps we can parlay it into something larger, as Searle tries to when he infers that mentality is 'intrinsic'.[4] Of several things that he means by this, the clearest is that attributions of mentality to things are absolutely true/false and not merely acceptable/unacceptable relative to the interests of the speaker. I agree with Searle about this, but it has nothing to do with constraints on inner routes from input to output. It also has little to do with any of my themes in this paper, so I shall not pursue it. Nor shall I discuss what else Searle means by 'intrinsic'.[5]

Dan Block has considered whether something like the S result might show that our concepts don't allow that a thinker could have a part that was also a thinker. He decided against this, on the basis of a still more flabbergasting thought experiment: suppose that the quarks composing a human brain are being separately manipulated by thinkers who have decided to make the quarks behave in accordance with what we fondly think are the laws of physics; don't you agree with Block that this is consistent with the attribution of beliefs and desires to the owner of the brain?[6] I have no answer to this. The story is so beset with noise that I can make nothing of it.

Anyway, why does it *matter* what boundaries on thinkers are set by our concepts? If for example our concepts don't allow that a thinker could be a proper part of a thinker, so what? A result along these lines would be significant only if it showed something about the central, active, operant part of our mentalistic conceptual scheme. If something in *that* won't let us classify anything both as a thinker and as a proper part of a thinker, that is interesting news. Here is the alternative:

> Our conceptual scheme has a busily active core that governs how we relate thoughts to one another and to environments and behavior. It also includes, sitting off to one side with no particular work to do in combination with the active core, a requirement that an item is not to count as a thinker unless it satisfies some condition C. Everything we know of that satisfies the core also satisfies C; the C-requirement is not something we use to divide up the world of actual prima facie thinkers into Passes and Fails.

If that situation obtains, the C-requirement is a wheel that turns although nothing turns with it; a negligible part of the mechanism. It's as though we had a concept of *human being* that included whatever our actual concept does include together with the requirement 'was not born with purple hair'. If we were foolish enough to have such a concept, that might be an interesting fact about our pathology, but it would not imply anything significant about conceptual structures.

It is unlikely that any part of our conceptual scheme does include anything like that—any large aspect to a concept that is logically detached from all its other aspects, not arising from the others or even combining with them to make itself felt in everyday thinking and talking. Why, after all, should our ancestors have conceptually forearmed themselves in this way? Indeed, if Quine is right about the difference between what is made true by our concepts and what is made true by how our concepts relate to the world, there cannot be a conceptual truth of the sort now in question. Quine holds that conceptual truth or analyticity, such as it is, results from the role the analytic sentence plays in our management of the interplay between some large class of sentences and the impingements of the world upon us. No such role, no analyticity!

If on the other hand our conceptual scheme does in a central, active way put some constraint C on all possible thinkers, e.g. some constraint on the inner route from their inputs to their outputs, there will be a better way of demonstrating this than by telling bizarre stories in which C is infringed and noting that readers aren't comfortable with them as stories about thinkers. The better way, unhindered by the noise problem, is to show how the core works and how the C-requirement arises out of those workings.

4. Causation and explanation

The most popular attempt to show along those lines that our concepts imply inner-route constraints goes like this:

> Our mentalistic conceptual scheme actively and centrally requires that we behave as we do *because* of our beliefs and desires. There is no non-magical way of making sense of this except by supposing that our beliefs and desires are among the causes of our conduct; this implies that beliefs and desires must be particular events or state-tokens or the like, because that is what causes are. So our conceptual scheme does make demands on the inner causal route from input to output, namely that it must run through particular items that can rightly be characterized as beliefs and desires.

This is wrong. Folk psychology does insist that attributions of beliefs and desires must help to *explain* behavior.[7] If you like, say that they must aid in *causally explaining* behavior. Indeed, go the extra step and say that facts about behavior are caused by facts about what creatures think and want, or that facts of the former kind are causal consequences of facts of the latter kind.[8] You still haven't implied that there is any such item as a belief, or as a desire. The explanatory and perhaps causal power of attributions of beliefs and desires does not require us—perhaps it does not even permit us—to reify beliefs and desires, treating them as countable particular items of some kind. For a simple analogy, think of the causal explanatoriness of statements about shortages ('There is a shortage of food in Ethiopia; there is no shortage of oil in Mexico'); such statements can have explanatory power without our reifying shortages, treating them as though they were particular items in the world—negative storage bins, perhaps.

The questions 'What *is* a belief?' and 'What *is* a desire?' need have no answers, then; and I believe that they have no answers. What can be answered are such questions as 'What kind of thing are we saying when we explain behavior by attributing beliefs and desires?'

A purely input/output account of intentionality may be complained of in the words 'It says things about when it is all right to attribute beliefs and desires, but it doesn't say what beliefs and desires are.' This could mean 'The account is purely externalist; it doesn't take us into the interior', or it could mean 'The account gives truth conditions for sentences using the verbs 'believe' and 'want' but doesn't give application-conditions for the nouns "belief" and "want"'.

Both complaints are misguided, I believe. But let us keep them apart: they are two complaints, not one.

What can we infer from the fact that attributions of cognitive mentality must be explanatory? In section 9 I shall give that question an answer that does not involve any inner-route constraints. But other things have to be done first.

5. The founding triangle

The concepts of belief and desire are linked to one another, and to behavior, in a famous triangle: an animal will *do* what it *thinks* will lead to what it *wants*. This does not have a very long ancestry: it can be found in Braithwaite's 'The Nature of Believing', which has a clear and acknowledged predecessor in Alexander Bain's work,[9] but I can't confidently run it further back than that. One might think of it as anticipated by what Hume says about beliefs as 'the governing principle of our actions' combined with his remark that reason is the slave of the passions, which might mean that cognitive states can affect behavior only when desires are also at work;[10] but no careful reader of Hume could think that he had the triangle clearly in focus, and insofar as he had it at all he derived it causally from a story whose conceptual foundations were entirely different.

Although the triangle thesis has won almost complete acceptance among philosophers in a little more than a century, we are not yet at the bottom of its implications.

This triangle is deeply teleological. I think that the best way to get an entry into intentionality—i.e. into the concepts of belief and desire—is through the idea of a system that *seeks a goal*, doing what it thinks will secure the goal. (I here rely on work that is presented more fully in my book *Linguistic Behaviour*, which develops ideas that were brought to a head in Charles Taylor's *The Explanation of Behaviour*.) Start with the suggestion that for x to have G as a long-term goal is for this to be true:

(1) Whenever x is so situated that it could get G by doing F, it does F.

If that is right, we can analyse 'x has G as a goal right now' or 'x wants G' along the lines of:

(2) x is now in a condition such that: for as long as x is in that condition **(1)** is true of it.

This is too simple in many different ways—e.g. what if x has more wants than it can fulfill? But only one inadequacy needs to be paraded here, namely the fact that **(1)** and **(2)** have no chance of being true except by accident. That x could get G by doing F is a fact about how x is situated, how it relates to various kinds of objects in its environment, and at our world no such fact can modify how x behaves. What does have a chance of affecting x's behavior is its *registering* the fact that it can get G by doing F, i.e. that fact's being somehow imprinted upon x. So what we need is to replace **(1)** by this:

(1') Whenever x registers that it could get G by doing F, it does F.

I have coined 'registration' to name a genus of which 'belief' names the most prominent species; the differentia doesn't matter just now.

In a very tight nutshell, then, the initial launching-pad for the notion of belief is a thing that conforms to a cognitive-teleological generalization of the type of **(1')**; and desire enters through this:

(2') x is now in a condition such that: whenever x is in that condition **(1')** is true of it.

This way of getting the belief-desire-behavior triangle into operation (and I know of no other) implies that teleology is at the foundations of cognitive mentality. Really, that is an almost trivial result for anyone who is convinced that cognitive mentality rests on a triangle of which desire is one of the sides. But it isn't enough just to declare that teleology is foundational in psychology—one needs an understanding of what teleology is that makes clear *how* it can be harnessed to concepts of cognition. One does not, for example, want the conceptual item that Ernest Nagel offered under the label of 'teleology', for the whole point of that was to keep mentality out. Wanting to show that there could be goal-pursuits without cognition, Nagel developed a concept of teleology that resists fusion with anything cognitive.[11] That is fairly typical of what happens in the philosophy of biology. A treatment of 'teleology' by William Wimsatt, for instance, is primarily an account of the notion of *biological function*, and considered as such it is impressive; but Wimsatt does not try to develop it into something that might lie at the basis of a philosophical treatment of cognitive psychology, and I do not see how he could succeed if he did try.[12] Yet William Lycan, in an influential paper emphasizing the importance of teleological foundations for psychology, instead of offering his own account of teleology or building on Taylor's, offers only a deferential wave in the direction of 'philosophers of biology', especially Popper and Wimsatt.[13] I protest that one needs to understand *how* teleological concepts might fit in with the rest, and such an understanding can be found not in the philosophers of biology but in the work of Taylor's that I have been developing. The emphasis on evolution that dominates the work on teleology by the biologically oriented writers poses an odd problem for anyone who believes, as Lycan and I do, that teleology is at the heart of our system of cognitive concepts. Of course actual teleology evolved: that true but irrelevant to a conceptual inquiry. There may be a sense of 'teleology' in which the existence of teleology conceptually requires evolution, but 'teleology' in that sense cannot be conceptually required for cognition. If it were, it would follow that it is absolutely, conceptually impossible that cognition should exist except as a result of evolution; presumably nobody believes a conclusion so fanciful. The one conceptual connection that *does* obtain between cognition and evolution will be presented in section 9.

6. The unalikeness of belief and desire

A grasp of how the cognitive teleology triangle actually works shows one something that is not grasped by those who only bow to the triangle from a distance—namely that belief and desire have almost nothing in common. Although facts about what an animal thinks and facts about what it wants collaborate to explain its behavior, the collaborators are enormously different from one another. Beliefs and desires are formally similar, in that each is a psychological propositional attitude, so that the very same propositional value of P could occur in 'x thinks that P' and 'x wants it to be the case that P'; but that is all the similarity there is, while the unlikenesses are many and striking.

For one thing, in the most fundamental story about the belief-desire-behavior triangle, the relevant beliefs are all about means to ends. The basic story goes like this (the language is a bit stilted, so as to keep the *formal* similarity in view):

> The animal did F because it believed that *doing F was a way to bring it about that P*, and it desired that P.

Here P is a dummy, which might be replaced by all sorts of things. In any unfanciful basic account, it will be a proposition about the animal's being in a certain state; in no unfanciful account will it be the proposition that something is a means to a certain end. In the ground-floor story, the animal's basic desires are never that x should be a means to y, whereas the basic beliefs are all about means to ends. Of course, a highly developed animal—such as you or me—could in principle believe anything at all (e.g. that he is going to be safe and warm) and could want anything at all (e.g. that eating ice-cream should reduce one's weight). But down in the simple, core situations it's not like that.[14]

For that reason, and perhaps for others, the relevant beliefs are likely to change rapidly, in lock-step with changes in the environment; desires can change, but are more likely to do so in response to internal changes (e.g. levels of satiety) than in response to changes in the environment.

The differences are so great that there seems to be no way of giving a Y-shaped analysis of the concepts of belief and desire, starting with their common features and then going on with what differentiates them. A satisfactory analysis must let them stand side by side, each on its own feet, collaborating but not overlapping to any significant degree.

Searle aims to do better than that.[15] He does not try to—and does not think one can—analyse the concepts of belief and desire in terms that don't involve 'intentionality', but he does offer a Y-shaped account of them, purporting to tell a substantive part of the belief-desire (or 'intentionality') story in a general way, before dropping down to the level of detail at which belief and desire are distinguished.

The generic part of Searle's account says that an intentional state is a mental state that can be satisfied; for each such particular state there is a proposition whose truth is needed and enough for the state to be satisfied. If right now animal x is in a state S that has proposition P as its condition of satisfaction, then at this moment S *attitudes* that P, where the 'attitude' is a blank verb, coined by me, that might be replaced by either 'believe' or 'desire'.

In the differentiating part of the account, desires are said to relate to their conditions of satisfaction differently from how beliefs relate to theirs. The difference is in 'direction of fit': where beliefs are concerned, the direction runs from world to mental state, with desires it runs the other way.

Searle's account of 'direction of fit' is not very crisp. He says that beliefs are 'supposed in some way to match an independently existing world' whereas imperatives are 'supposed to bring about changes in the world' (p. 7), and speaks of where the 'fault' or the 'responsibility' lies if fit is not achieved. He could improve upon these formulations, I think, by saying that necessarily we try to make our beliefs fit the world and necessarily we try to make the world fit our desires. This could not be part of an analysis of intentional notions in non-intentional terms, because it uses 'try to', which has the concept of wanting buried in it. Searle is not looking for an analysis, however, so perhaps he is entitled to the differentiating part of his account.

That won't do much good, however, unless the generic story about beliefs and desires as states having conditions of satisfaction is all right. Is it? Well, 'satisfaction' is almost Searle's favorite technical term, yet he never explains it and it does not occur in his Index. So far as I can discover, our only grip on his notion of satisfaction comes through the double thought: if P then **(i)** the desire that P be the case is 'satisfied' in a normal sense of that word, and **(ii)** the belief that P is 'satisfied' in the special sense of being true. The disjunctive nature of this is evidence that we don't have any unitary concept of satisfaction that does Searle's work. From that I infer that he has not succeeded in giving a useful genus-and-then-species account of belief and desire. I doubt that such account could be given.

7. The unity condition

The triangle generates another line of thought, to which I now turn. The triangular conceptual structure is illustrated by the behavior of a thermostat: the thermostat 'wants' the room to be warmer, 'thinks' that closing the switch will bring this about, and accordingly closes the switch. But this illustration, though it is instructive, is also dangerous, as I shall now explain.

All the behavior of the thermostat that might be handled teleologically, or in intentional terms, is explained by a single mechanism, a single kind of causal chain that can be fully described without any use of intentional concepts. We can replace 'The thermostat does what it can to keep the temperature of the room close to 68 degrees' by 'The thermostat's switch closes whenever its temperature falls to 66 degrees and opens whenever its temperature rises to 70 degrees', and we can explain the latter generalization without any mention of 68 degrees as a goal and without mentioning beliefs and desires or anything like them.

In short, the one intentional account of the thermostat's behavior is matched by a single mechanistic account; and I submit that when that is the case, the latter account should prevail and the former, though perhaps stimulating and interesting for philosophical purposes, is false and should be rejected. For genuine teleology or intentionality, I contend, *the unity condition* must be satisfied. That is, a system x's intentionality is genuine only if

> Some class of x's inputs/outputs falls under a single intentional account—
>
> > involving a single goal-kind G such that x behaved on those occasions because on each of them it thought that what it was doing was the way to get G
>
> —and does not fall under any one mechanistic generalization.

Where that is satisfied, applying intentional concepts to the system brings a conceptual *unity* to some set of facts about it—a set that is not unifiable under a mechanistic description.

The unity condition marks off the systems some of whose behavior falls into intentional patterns that are not coextensive with mechanistic patterns. Only if a system's behavior satisfies that condition, I contend, is it legitimate for us to exploit its intentional patterns in our thought and speech. The marking-off is of course a matter of degree. It rejects intentionality when the intentional pattern coincides with a single mechanistic one; it welcomes it when such a pattern utilizes thousands of different mechanisms; and for many intermediate cases it gives an intervening judgment: 'Intentionality in this case is so-so—permissible but not very good.'

The fuzzy line drawn by the unity condition matches a lot of our intuitive sense of what systems do and what ones don't have thoughts and wants. Consider a chameleon flicking out its tongue and catching a fly with it. One can plausibly think of this as goal-pursuing behavior: it wants to eat the fly and thinks that this is the way to bring that about. But suppose we find that one uniform physical mechanism controls this pattern of behavior—a relatively simple causal tie between proximity of fly and movement of tongue, and between location of fly and direction of tongue movement, with, in each case, a few parameters in the one governing a few parameters in the other. Thoughtful people will regard this as evidence that the cognitive-teleological account of the behavior was wrong because really only a single mechanism was involved. The plausibility of the response 'Oh, so *that*'s all it was' is evidence for the truth of the unity thesis.

But we don't have to rely on such intuitive evidence. The unity thesis also corresponds to the best *defense* there is for using intentional concepts. The question of the legitimacy of intentional explanations of behavior ought to be faced squarely. Since chemical explanations involve principles that go wider and deeper, and theoretically admit of greater precision, why should they not always be preferred to explanations in terms of thoughts and wants?

Well, they would not be preferable if there were animal movements that could not be explained chemically but could be explained in terms of thoughts and wants. But none of us thinks that that ever actually happens. Again, explanations in terms of cognitive teleology might be adopted because we didn't know what the mechanistic, efficient-cause explanations were; but I hope we can do better than that. The remaining possibility is the one yielded by the unity thesis, namely that an explanation in intentional terms might be justified because it brings out patterns, provides groupings and comparisons, which a chemical explanation would miss. What the animal did belongs to a class of behaviors in which it wants food and does what it thinks will provide food, and there is no unitary chemical explanation that covers just this range of data. This animal seeks food in many different ways, triggered by different sensory inputs, and it is not credible that a mechanistic, physiological view of the facts will reveal any unity in them that they don't share with behaviors that were not food-seeking at all. If this unifying view of the facts answers to our interests, gives us one kind of understanding of the animal, and facilitates predictions of a kind that are otherwise impossible (predictions like 'It will go after that rabbit somehow'), we have reason for adopting it. These reasons leave us free still to acknowledge that each of the explained facts, taken separately, admits of an explanation that is deeper and more wide-ranging and—other things being equal—preferable.[16]

8. Morgan's Canon

What is it?

It is often held by philosophers of mind that there are senses of 'higher' and 'lower' that make true something that I shall, nearly following Dennett, call Morgan's Canon:

> In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty if it can be interpreted as the outcome of the exercise of one that stands lower in the psychological scale.[17]

This and its ilk have been praised and obeyed by many of us, without attending closely enough to what it says or to why it is true. I shall try to get clearer about both.

To make any progress with this, we have to realize that the Canon is useless when applied to the interpretation of particular behavioral episodes taken separately. Every individual bit of behavior—every simple or complex animal movement—can be interpreted as the outcome of chemical goings-on or (a little higher up the scale) of a virtually mindless stimulus-response mechanism. Whether such an interpretation is correct can be answered only by trying it on classes of behavioral episodes. So I take it that the Canon should be understood to be saying:

> In no case may we interpret a class of actions as the outcome of the exercise of a higher psychical faculty if they can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale.

Furthermore, I take it that whatever 'the psychological scale' is, anything that is on it is 'higher' than anything that is off it. That makes the Canon imply that a mentalistic explanation of a class of behaviors is wrong if the behaviors could be given a unitary mechanistic explanation, such as a chemical one. With the Canon thus interpreted, it turns out to entail my unity thesis. What else does it imply? That is, of the items that *are* on the psychological scale, what is higher than what?

Well, presumably explanations in terms of stimulus-response patterns, whether hard-wired or learned, are as low as you can get on the scale, so that Morgan's Canon implies that a class of behaviors that can be given a single stimulus-response explanation ought not to be explained in terms of beliefs and desires or anything like them.

As between two explanations that both attribute beliefs and desires, what could lead us to suppose that one is higher than another? Relative logical strength isn't a help. 'P entails Q' isn't necessary for 'P is higher than Q', we must hope: Morgan's Canon wouldn't amount to much if it could never be applied between attributions that were logically unrelated, as in 'The animal thinks there is a cat in the tree' and 'The animal thinks that predators are less likely to be around when there has been rain within the past three days or snow within the past two'. Nor is 'P entails Q' sufficient for 'P is higher than Q': it would be a funny psychological scale that put 'It thinks that the cat is up the tree' higher than 'It thinks that either the cat is up the tree or predators are less likely to be around when there has been rain within the past three days or snow within the past two'.

Christopher Peacocke, who is one of the few philosophers to have attended much to this matter, offers what amounts to this criterion: Attribution S is lower (he says 'tighter') than S* if every concept involved in S is involved in S*, and not vice versa.[18] It follows, as Peacocke points out, that a tightness comparison is possible only if one 'family of concepts has strictly greater expressive power than the other'. But that does not mean that a comparison of tightness can be made only when one of the two attributions entails the other: the relation of entailment or inclusion is not between propositions but between sets of concepts involved in propositions.

The concept-inclusion criterion is relevant to issues like this: Should the animal's behavior be explained in terms of what it thinks about its present environment, or rather in terms of **(i)** what it thinks about some earlier state of affairs and **(ii)** what it thinks about how such states affairs develop through time? Does it dig in that manner just because it thinks there is a bone under there, or does it do so because it thinks **(i)** that the bone was there yesterday and **(ii)** that buried bones generally stay put? The second diagnosis attributes to the animal a concept of the past, and a generalizing concept, and because of these it counts as higher than the other.

There are many cases where the concept-inclusion criterion needs help if it is to yield the answer we intuitively want. It seems right to suppose that 'The animal wants the others to stop eating' is lower than 'The animal wants the others to think that there is a predator in the vicinity'; but it is not clear that the former is 'tighter' by Peacocke's criterion, because each statement attributes a concept not attributed by the other. It might be replied that the concept-inclusion criterion involves not only the concepts that are directly attributed but also ones that the animal would have to possess in order to have the ones that are directly attributed. That might seem to deal with the case in hand, if we suppose that no animal could have the concept of another animal's *thinking that P* for any value of P unless it also had the concept of another animal's *doing* A, for various values of A. But what is needed is something stronger and less plausible, namely that no animal could have the concept of another animal's thinking that P for any value of P unless it also had the concept of another animal's *eating*.

Anyway, concept-inclusion is clearly irrelevant to some clear cases of level-difference. Consider the choice between

'The animal wants the others to think there is a snake in the undergrowth' and 'The animal wants the others to think that she thinks that there is a snake in the undergrowth'. These use exactly the same conceptual repertoire, but intuitively one would want one of them to count as 'lower' than the other: that is, one wants to understand Morgan's Canon in such a way that it condemns the explanation 'When she sees a snake she gives that cry because she wants the others to think that she thinks that there is a snake in the undergrowth' if the behavior in question could as well be explained by 'When she sees a snake she gives that cry because she wants the others to think that there is a snake in the undergrowth'. But the difference here is not in what concepts are involved, but in the level of iterative complexity with which they are involved.

That is the best I can do to mark out the steps in the 'psychological scale': from non-mentalistic to stimulus-response to belief-and-desire; and then from poorer to richer stocks of concepts, and from lesser to greater degrees of iterative complexity. The only way I can find of generalizing across the entire scale is to say that if P is lower than Q just in case it attributes to the animal less complexity of structure than Q does, or that the attribution Q implicitly credits the animal with every capacity credited to it by P and some more as well. This is no triumph of hard-edged technical clarity; but it has some content, is not intolerably fuzzy, and draws the line in a plausible place.

Where and why is the Canon binding?

It is obvious why we need some constraints on the attribution of mentality to animals. To put it mildly, the results of sloppy, easy, unconstrained attributions are boring because contentless. Conceptual analysis tells us that much; and I have contended that it also tells us that the unity condition should constrain our explanations of behavior, that is, that we shouldn't explain in terms of beliefs and desires a class of behaviors that could be given a unitary mechanistic explanation. But what about the rest of Morgan's Canon? Granted that within the domain of belief-desire explanations we need some constraints, what is the warrant for saying that the constraints must have the effect of always pushing explanations as far down the 'psychological scale' as possible?

If there is one, it is presumably a warrant for saying that we should in general assume things to be homogeneous or unstructured except to the extent that the evidence points to structure or complexity. That seems to be what Peacocke has in mind in his brief statement of why the 'tightness condition' is a valid constraint: 'Without this requirement, the attribution of concepts is unconstrained by the presence of intentional actions responsive to the distinctions in the world drawn by these concepts' (p. 85). But so what? Neither Peacocke nor I has yet given a clear, convincing statement of what intellectual sin is being committed by someone who infringes the Canon on this understanding of it. In considering this, we have to look at two different kinds of situation—two kinds of rivalry between competing explanations of a class of behaviors.

In one of them, each of the rival explanations accords with the facts so far observed, but they differ in their implications for behavior and are thus not empirically equivalent. In that case, the main advice to be given, having nothing to do with Morgan's Canon, is: Look for further data that fit one of them and not the other, perhaps by setting up situations designed to elicit behavior that will serve as an *experimentum crucis*.[19] For example, in trying to adjudicate between 'When the animal screams like that it is because it wants its companions to climb trees' and 'When the animal screams like that it is because it wants its companions to think there is a leopard in the vicinity', there is no theoretical need for Morgan's Canon. Each of those hypotheses, unless specially padded with supplementary hypotheses of a sort I'll discuss shortly, has behavioral implications that the other lacks, and the final arbiter should be the behavior of the animal in crucial situations.

If the Canon has work to do here, it is only as advice about what we should provisionally believe while waiting for the issue to be settled empirically—advising us that in the meantime it would be wise to expect the decisive data to rule out the 'higher' hypothesis and thus to favor the 'lower' one. It could also be advising us about what we should tentatively believe if it's now too late ever to get the question settled, e.g. because the animal is dead and its species extinct.

This is good advice on our planet, where most mentality is fairly low-level. But there could be planets where most of the known minds were high-level, sophisticated ones, and where most teleological patterns in animal behavior were not reducible to mechanistic ones. On such planets the provisional advice issued by Morgan's Canon would be bad. So this use of the Canon reveals nothing about our mentalistic conceptual scheme.

The other kind of rivalry is that between two explanatory hypotheses which, though they differ in the 'height' of the psychological capacities that they attribute, are empirically equivalent.[20] This can happen only if the 'higher' one includes supplementary hypotheses to explain why the extra psychological capacity is not manifested in behavior. For example, the lower one might be:

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

while its rival says that

> The animal has the concepts one, two, three, four, and equal-numberedness, but it can't be got to use its concept of four in any way except in doing number comparisons between quartets and other groups.

The second of these says not that the animal never has a reason to use *four* for anything except comparing quartets with other groups, but rather that even in situations where some other use of *four* would be to the animal's advantage, it reliably doesn't use its concept of four in that other manner.

How are we to choose between these? Well, one of them credits the animal with two more items than the other does—namely an extra concept, and an impediment to its being implemented in most situations. So far as I can see, all we need here is a quite general principle that should regulate us in theory building, namely: Prefer what is simple to what is complex, unless there is independent justification for the complexity. What could justify the complexity? It would have to be something that I don't want to discuss in this paper, namely a theory of the animal's internal cognitive dynamics—that is, that part of our psychological account of it that speaks of how changes in its beliefs cause not only behavior but other changes in its beliefs, and so on. Suppose we are constructing such a theory for an animal, and are faced with the rival stories about its grasp of the number four, one possibility is this: Our smoothest and generally most plausible explanation for the animal's grasp of one, two and three implies that it does also have the concept of four; and we have good evidence for its having a natural class of cognitive obstacles that would include an inability to employ *four* except in number comparisons between quartets and other groups. In that (admittedly fanciful) case, we might justifiably prefer the more complex hypothesis and thus, incidentally, prefer the 'higher' to the 'lower' psychological attribution to the animal. Without something like that, the 'lower' should be preferred, not because it is 'lower' but because it is less complex and greater complexity is not justified. This result coincides with what Morgan's Canon would say, but the Canon has nothing to do with it.

Considered as a rule of thumb to guide our provisional opinions about cognitive abilities, Morgan's Canon is fine. Considered as anything else, it is negligible. Having worshipped at its shrine for a quarter of a century, I say this with regret.

9. Intentionality as a source of explanations

The concepts of belief and desire are fundamentally explanatory. In the account I have been giving, explanatoriness is supposed to come in through the generalizations that define an animal's goals—i.e. ones of the form 'Whenever x registers that it could get G by doing F, it does F'. But for all I have said to the contrary, such a generalization might be true merely by coincidence, which would unfit it to explain anything. That is, it might be a mere coincidence that this single system houses a lot of mechanisms whose over-all effect is to make the system a G-seeker; and if it is a coincidence, the system's intentionality cannot be used to explain its behavior. (By 'mechanism' I mean 'physical feature that makes it the case, for some value of I and some value of O, that if the animal receives input of sensory kind I it produces behavioral output of motor kind O'.)

In plugging this gap in the account, I shall exploit the link between what *can explain* and what *could have predicted*. That is, I shall look for conditions under which a teleological generalization could be used to explain an animal's moving in a certain way at a particular time by looking for the conditions under which the generalization could be used to predict that the animal would move like that then. For this to serve my purpose, we have to be able to predict a link between one sensory kind of input and one motor kind of output on the basis of links between other pairs—ones in which the sensory kinds are different (and perhaps the motor kinds as well). That is, I want to know what can entitle us, when we know that an animal goes after rabbits in many different ways on the basis of many different sensory kinds of clue, to take that as *some* evidence that it will go after rabbits on the basis of kinds of clue that we haven't so far observed it to use.

There seem to be just two ways of supplying this need.

One of them uses evolutionary ideas. Simply and abstractly: of all the potential mechanisms that got an initial genetic hold on the animal's ancestors through random mutations, relatively few survived; among the survivors were the bunch of mechanisms that make their owner a G-getter, and *that is why they survived*. Why does this animal contain a lot of mechanisms that make it a G-getter? It inherited those mechanisms from a gene pool that contained them *because they are mechanisms that make their owner a G-getter*.

That makes it more than a coincidence that the animal has many mechanisms that are united in their G-getting tendency, and lays a clear basis for explanations that bring in cognition. That a species has evolved a G-getting tendency that is manifested in this, that and the other links between sensory kinds of input and motor kinds of output creates some presumption that it has evolved other links that also have a G-getting tendency. So there is something predictive in this, and thus something explanatory as well.

(An analogous story could be told, without evolution, if animals were made by a creator who intended them to manifest the patterns of cognitive teleology. The animal does G-getting things on receipt of clues of various types; that is some evidence that its designer wanted it to be a G-getter, which is some reason to think the animal will do G-getting things when it has clues of other kinds that we haven't yet seen it respond to. This would raise the question of how we are to understand statements about what the creator thinks and wants, and one might wonder whether *that* could be tackled along the functionalist lines that have informed this paper. I shall not pursue the matter.[21])

Notice that the evolutionary source of explanatoriness does not require that the animal be educable, flexible, capable of adapting as an individual to what it learns about its world.[22] The account would go through quite well for an animal like this: It picks up from its environments all kinds of information about ways to get G, and acts accordingly, but if one of these input-output pairs starts to let it down, leading not to G but to something unpleasant, that does not lead the animal to delete that input-output pair from its repertoire. Nor does it ever add anything to its repertoire in the light of chance discoveries about what works.

It is vastly improbable that any species should evolve the kind and degree of complexity that satisfies the 'unity thesis' without also evolving a degree of individual adaptability to what experience teaches. But we know at what sort of world (or planet) such a thing might occur in the course of nature: it would be a world where behavioral complexity had great survival value whereas individual adaptability didn't. And the supposition itself clearly makes sense: it is the supposition of behaviorally frozen animals with a behavioral repertoire that falls into teleological patterns that don't map onto patterns of any other kind. Such animals would cope successfully and (it would seem) intelligently with their environments, but as soon as these altered a bit in some relevant way, the animals would be stuck.

To repeat what I said a moment ago, the behavior of such creatures could be explained and predicted through the generalizations of cognitive teleology. If an animal has a lot of (for short) G-seeking input-output patterns, that is evidence that they have been selected *because* they let the animal get G; and *that* is evidence that other input-output links that have the same upshot will also have been selected. By the prediction test, therefore, we can use the premise that the animal is a G-seeker to explain a new bit of G-seeking by it; the premise is at least somewhat projectible, and is not a mere summation of observed behavioral episodes. In short: evolution could make cognitive mentality explanatory, even if the animal could not learn from its experience.

I emphasize this in order to introduce the point that individual educability can make cognitive mentality explanatory even if the animal had not evolved.

Consider the case of educable parents that have an educable offspring with a goal that they didn't have: the offspring is the locus of a large number of G-getting mechanisms, none of which were present in the parents, their presence in the offspring being the result of a very radical and sheerly coincidental set of genetic mutations. This story, though utterly improbable, states a real conceptual possibility; and if we knew that it was true of a given animal, we could *explain* some of the animal's behavior in terms of its having G as a goal. For **(i)** its having G as a goal and **(ii)** its being able to learn from experience jointly give us reason to predict that it will pursue G in ways (and on clues) that we have not previously seen it employ (assuming that the animal may indeed have previously employed those ways and clues or ones that were suitably related to them).

So our conceptual scheme does not insist that believers and wanters must have evolution (or a personal creator) in their causal ancestry. The scheme does demand that attributions of belief and desire be capable of supporting explanations of behavior; this requires a context where the generalizations of cognitive teleology license predictions; and two such contexts suffice for this—evolution and individual educability, or Mother Nature and soft wiring. Please note that I have not invited you to consider intuitively whether you would be willing to attribute beliefs and desires to an educable animal that didn't evolve (and wasn't personally created), or to an evolved animal that wasn't educable. I have tried to hush all that noise, and show what follows from something that can be found at the humdrum, familiar, smoothly-running core of our cognitive conceptual scheme.

Notes

[1] W. V. Quine, 'Two Dogmas of Empiricism', in his *From a Logical Point of View* (Harvard University Press: Cambridge, Mass., 1953), pp. 20–46. For an account of how this contends that analytic/synthetic is a difference of degree, see the first three sections of Jonathan Bennett, 'Analytic-Synthetic', *Proceedings of the Aristotelian Society*, new series vol. 59 (1958–9), pp. 166–188. The fourth section is no good.

[2] That is why it is right for Lewis to base an analysis on 'platitudes. . . regarding the causal relations of mental states, sensory stimuli, and motor responses'. He is interested only in 'platitudes which are common knowledge among us—everyone knows them, everyone knows that everyone knows them, and so on', because 'the meanings of our words are common knowledge, and I am going to claim that the names of mental states derive their meanings from these platitudes'. David Lewis, 'Psychophysical and Theoretical Identifications', *Australasian Journal of Philosophy* 50 (1972), pp. 249–258, at p. 256.

[3] John R. Searle, 'Minds, Brains, and Programs', *The Behavioral and Brain Sciences* 3 (1980), pp. 417–424; 'Analytic Philosophy and Mental Phenomena', *Midwest Studies in Philosophy* 6 (1981), pp. 405–424

[4] Searle's thesis that only animals can think is supposed to be based (I don't understand how) on what we know about actual animal thinkers rather than on what we intuitively think about system S. (Searle, 'Analytic Philosophy and Mental Phenomena', *op. cit.*, at pp. 413f.)

[5] For a discussion of them, see Paul M. Churchland and Patricia Smith Churchland, 'Functionalism, Qualia, and Intentionality', *Philosophical Topics* 12 (1981), pp. 121–145, at pp. 139–141.

[6] Ned Block, 'Troubles with Functionalism', in N. Block (ed.), *Readings in Philosophical Psychology*, vol. 1 (Harvard University Press: Cambridge, MA, 1980), pp. 268–305, at pp. 279f.

[7] Evidence for this comes from the failure of attempts to explain the relevant concepts in ways that make them primarily descriptive and only secondarily and inessentially explanatory. See Jonathan Bennett, *Linguistic Behaviour* (Cambridge University Press: Cambridge, 1977; Hackett: Indianapolis, 1989), section 13.

[8] For a defence of the concept of fact causation, see Jonathan Bennett, *Events and their Names* (Hackett: Indianapolis, 1988), chapter 2.

[9] R. B. Braithwaite, 'The Nature of Believing', first published in 1932 and reprinted in A. P. Griffiths (ed.), *Knowledge and Belief* (Oxford University Press: Oxford, 1967). Alexander Bain, *The Emotions and the Will*, first published in 1859 and re-issued by University Publications of America: Washington, D.C., 1977; the relevant part is the chapter entitled 'Belief', especially pp. 568–573 in the reprint.

[10] David Hume, Hume, *An Enquiry Concerning Human Understanding*, pp. 49f in the Selby-Bigge edition; *A Treatise of Human Nature* II.iii.3 (pp. 413–8 in the Selby-Bigge edition).

[11] Ernest Nagel, 'Teleology Revisited', in *Teleology and Other Essays* (New York, 1979), pp. 275–316.

[12] William C. Wimsatt, 'Teleology and the Logical Structure of Function Statements', *Studies in History and Philosophy of Science* 3 (1972).

[13] William Lycan, 'Form, Function, and Feel', *Journal of Philosophy* 78 (1981), pp. 24–50, at p. 32. In Lycan's more recent *Consciousness* (M.I.T. Press: Cambridge, Mass., 1987) at p. 43 the deferential wave occurs again, still in the direction of 'philosophers of biology', with special emphasis this time on unpublished work by Karen Neander.

[14] What gives us our unlimited repertoire of possible beliefs and desires is, I think, our ability to articulate them in language. See Jonathan Bennett, *Rationality* (Routledge and Kegan Paul: London, 1964; Hackett: Indianapolis, 1989); 'Thoughtful Brutes', *Proceedings of the American Philosophical Association* 62(1988), pp. 197–210.

[15] John R. Searle, *Intentionality* (Cambridge University Press: Cambridge, 1983), ch. 1.

[16] For more along this line, see Jonathan Bennett, *Linguistic Behaviour*, *op. cit.*, sections 21–22; Daniel C. Dennett, *The Intentional Stance* (M.I.T. Press: Cambridge, MA, 1987), chapter 2.

[17] C. Lloyd Morgan, *Introduction to Comparative Psychology* (London, 1894), p. 53. Quoted from Christopher Peacocke, *Sense and Content* (Oxford University Press: Oxford, 1983), p. 86n. Dennett calls it 'Lloyd Morgan's Canon', but so far as I can discover the author's surname was simply 'Morgan'.

[18] Christopher Peacocke, *Sense and Content: Experience, Thought, and their Relations* (Clarendon Press: Oxford, 1983), p. 84. Peacocke's discussion of Morgan's Canon (pp. 78–86) in is the spirit of my *Linguistic Behaviour*, *op. cit.*, sections 36 and 37.

[19] My unity thesis can be developed into some suggestions about what sort of evidence would adjudicate between rival explanations. For details, see my 'Folk Psychological Explanations', in John D. Greenwood (ed.), *The Future of Folk Psychology: Intentionality and Cognitive Science* (Oxford University Press, 1989).

[20] Rivalries between empirically equivalent hypotheses that differ in what they attribute, but not in the 'height' of what they attribute, are beside my present point because they don't invite us to invoke Morgan's Canon.

[21] For a determined assault on the problem of giving a functionalist analysis of attributions of cognitive states to a deity, see William P. Alston, 'Functionalism and Theological Language', *American Philosophical Quarterly* 22 (1985), pp. 221–30; 'Divine and Human Action', in T.V. Morris (ed.), *Divine and Human Action* (Cornell University Press: Ithaca, NY, 1988), pp. 257–80.

[22] Nor does the explanation in terms of a personal creator. From now on I shall forget the personal-creator option, and focus on the other. This is just for simplicity's sake. I justify it on the grounds that it is quite certain that actual animals did evolve. Perhaps a personal creator somehow made them do so; but as long as evolution by natural selection *did* occur, that is enough to give predictive and explanatory force to our attributions of beliefs and desires, whether or not the forces of evolution are themselves an expression of divine intent.