Mentalism and the Explanation of Behaviour 3

Mentalist Explanations - Believing and Wanting

The Nature of Mentalist Explanation

A *mentalist explanation* in the sense in which I propose to use that term for our present purposes may be defined as the attempt to predict or explain *ex post facto* the behaviour of an individual human being or other living organism in terms of (a) what he *knows* or *believes* about the situation confronting him and the probable consequences of the different courses of action open to him, and (b) the various contingencies which he *wants* either to bring about or prevent.

The scope of mentalist explanation

Mentalist explanations in this sense are designed to explain and predict particular actions of the individual concerned rather than recurrent behavioural phenomena. They can sometimes provide a partial explanation of certain recurrent behavioural phenomena which are peculiar to a particular individual such as his habits or his areas of emotional sensitivity; but they cannot properly be used to account either for traits of character and personality peculiar to the individual or for behavioural phenomena exhibited by all members of a given species or across different species of organism.

Mentalist explanations are also restricted in their application to different types of particular actions. In the last lecture I distinguished four different varieties of human action which are recognised within the framework of the psychological concepts of ordinary language:

1. Deliberate or premeditated actions based on a preformed intention to act
2. Voluntary actions which are not deliberate in this sense amongst which we can distinguish two sub-varieties:
   (a) actions done from force of habit
   (b) actions done on impulse.
3. Unconscious, involuntary, yet purposive actions such as hysteria
4. Involuntary reflex actions

In terms of this classification, we can say that all these varieties of action except the last one - involuntary reflex action - are at least partly, susceptible to explanation and prediction in mentalist terms. Moreover, as we saw last time, there is a convention in ordinary language whereby the agency in the case of involuntary reflex action is attributed to the part of the body involved rather than to the individual as a whole, which would allow us to say that everything a man does (as opposed to what his limbs and organs do) is susceptible to mentalist explanation. However in the case of habitual action all that the mentalist explanation can do is to explain why the individual came to adopt this course of action in the first place. Any prediction of the occurrence of the behaviour in advance is based solely on the observation that it has repeatedly occurred under similar circumstances in the past. Moreover, in the case of unconscious and involuntary purposive behaviour, although the hysteric may quite properly be said to act from desires and fears of which he is unconscious, there are serious conceptual objections to describing him as acting on an unconscious belief or an unconscious intention. It is only in the case of deliberate premeditated action and unpremeditated action on impulse that mentalist explanations can provide an account which is both entirely appropriate and reasonably comprehensive.

The dispositional character of mentalist explanations

As we saw in Lecture 6-1 we usually explain particular occurrences or events in terms of the occurrence of a particular triggering event in conjunction with the dispositional properties of the substance or substances involved. Now according to Ryle (5) to whose view I subscribe in this respect, the explanatory concepts 'knowing', 'believing' and 'wanting', whose employment constitutes the defining characteristic of a mentalist explanation in the sense in which we are using that term here, are all dispositional concepts. If this

view is correct, it follows that when we explain why someone did something in terms of what he knew or believed and what he wanted to achieve thereby, we are explaining the behavioural event in question in much the same way as we would do, if we attributed the fact that a particular piece of glass broke on a particular occasion to its brittleness or the speed with which the car climbed the hill to the horse power of the engine.

There are, however, certain important logical differences between concepts like 'knowing', 'believing' and 'wanting' and concepts like 'brittleness', 'flexibility', 'being magnetized' and 'horse power' which have been used as paradigm cases of dispositional concepts in the context of the controversy which we examined in Lecture 6-1. In the first place verbs like 'knowing', 'believing', 'wanting' and 'intending' do not by themselves specify any particular dispositional property of the individual who is said to know, believe, want or intend. Indeed, to say that someone knows, believes, wants or intends says precisely nothing about him until we are told what it is that he knows, believes, wants or intends. Furthermore, the behaviour to be expected of someone who knows, believes, wants or intends one thing will be quite different from the behaviour to be expected of someone who knows, believes, wants or intends something else. In other words each item of knowledge, each belief, each desire or intention constitutes a distinct and separate dispositional property. The same is true of other dispositional concepts of a psychological or behavioural kind such as the concept of 'habit'. Here again, to say that someone has a habit tells us nothing, until we are told what it is he has a habit of doing. On the other hand there are dispositional concepts in the mental or psychological domain, such as the concept of 'being irritable', which *do* specify the type of behaviour to be expected of the individual without the need to specify the particular object of the mental disposition in question.

There is however, a further difference which distinguishes the concepts of 'believing' 'knowing' and 'wanting' employed in mentalist explanations from otherwise similar dispositional concepts like 'intending to do' or 'being in the habit of doing something' in that to say that someone knows something, believes something or wants something is not by itself sufficient to specify the kind of behaviour to be expected of the individual concerned. As Geach (2) points out in Chapter 4 of *Mental Acts* an action cannot "be described as 'acting as if you held such-and-such a belief' unless we take for granted, or are somehow specially informed about, the needs and wants of the agent." By the same token, we may add that, except in the special case where a man is described as wanting *to do* something, we cannot describe an action as a case of 'acting as if you wanted or needed something' unless we take for granted, or are somehow specially informed, that you know or believe that the action in question is likely to bring about the object of your desire. This feature of mentalist explanations of behaviour whereby the behaviour to be expected cannot be predicted from a knowledge of what a man believes, if you don't know what he wants, or from a knowledge of what he wants, if you don't know what he believes, is not a feature that is peculiar to dispositional concepts within the mental or psychological universe of discourse. A similar situation applies in the case of the dispositional concepts of potential difference and resistance which combine, according to the relationship expressed in Ohm's Law, to determine the flow of an electric current through a conductor. Here as in the case of beliefs and wants, the outcome cannot be predicted from a knowledge of the potential differences between the two ends of the conductor or from a knowledge of its resistance taken separately, but only from a knowledge of both in combination.

The intentionality of mental dispositions

Although, as we saw when we discussed the problem of intentionality in Lecture 8, none of the definitions of intentionality which have been put forward so far will allow us to maintain that all the psychological verbs in ordinary language take objects which are intentional in the sense of the definition in question, there can be no doubt that it is a distinctive feature of those dispositional concepts which occur in mentalist explanations that they are expressed by verbs which take what, by Geach's non-Shakespeariansity criterion (2) are intentional objects. This at least, is a feature which has no parallels in the case of non-mental dispositional concepts.

Nevertheless there is, in my view a close conceptual connection between the dispositional character of concepts like 'knowing', 'believing', 'wanting' and 'intending' and the fact that, *qua* verbs, they take non-Shakespearian and hence intentional objects. It is true that these mental disposition verbs are not the only psychological verbs in ordinary language which take intentional objects by Geach's criterion. But if we

examine those which do take objects of this kind, we shall find that although their primary reference is to mental acts and mental activities of various kinds rather than to dispositions, the propositions which contain them invariably entail either the antecedent, concurrent or consequent existence of some knowledge, belief, desire or intention of the individual concerned and it is the intentionality of the objects of the knowledge, belief, desire or intention which is entailed by these expressions, I suggest which accounts for the intentionality of the object of these mental acts and mental activity verbs.

Thus mental act verbs like 'recognise' and 'perceive' take intentional objects, because and in so far as they entail a subsequent and consequent 'knowing that p'. Mental act verbs like 'judge', 'conclude' or 'infer' take intentional objects because and in so far as they entail a subsequent and consequent 'believing that p.' Similarly the mental act verb 'decide' takes an intentional object because and in so far as it entails a subsequent and consequent 'intending to Φ'. The mental activity verb 'trying' takes an intentional object because and in so far as it entails an antecedent and concurrent 'intending to Φ'. The mental activity verb 'looking' or 'searching' takes an intentional object because and in so far as it entails 'wanting to find something'. The only expressions, taking intentional objects by Geach's non-Shakespearianity criterion, which do not perhaps fit quite so neatly into this pattern are mental act expressions like 'It occurred to A that possibly p', 'A wondered if p' and 'A dreamed that p' which entail something like 'A was momentarily tempted to believe that p' rather than 'A believed that p' and mental activity expressions like 'thinking' or 'dreaming about O' and 'imagining what it would be like if p' which likewise do not imply commitment to any particular beliefs. It is arguable however, that one can only be said to entertain a proposition which one does not subsequently believe in so far as one knows what it implies and hence, what would be involved in believing it to be true.

If in the light of these considerations, we are justified in viewing the problem of intentionality as a problem about the objects taken by these four mental disposition verbs 'know', 'believe', 'want' and 'intend', it becomes apparent that intentional object verbs are of two types (a) the *verbs of cognition,* i.e. 'know' and 'believe' and mental act and activity verbs which entail 'knowing' and 'believing' and (b) the *verbs of volition* i.e. 'want' and 'intend' and the mental act and activity verbs which entail 'wanting' and 'intending'. Moreover once we make this classification it becomes apparent that there is an important difference between the kind of intentional object involved in the two cases. For the object in the case of a verb of cognition is always a *proposition.* whereas in the case of a verb of volition the intentional object is something that answers to a particular *description* or falls under a particular *concept.* We can express this difference by saying that verbs of cognition describe *propositional attitudes*, whereas verbs of volition describe what, in terms of the concept of *schema* which we discussed in the previous lecture, we may call *schematic attitudes.* When we characterise an individual's behavioural dispositions in terms of his propositional and schematic attitudes, what we are doing in both cases I suggest, is characterising his behavioural propensities in terms of how he is inclined to talk about the situation confronting him.

As Geach (2) has again pointed out in Chapter 4 of *Mental Acts*, we characterise what a man knows or believes "by using the *oratio obliqua* construction - i.e. the same construction as is used with 'verbs of saying' to report the gist or upshot of somebody's remark rather than the actual words he used." This observation has two important consequences. In the first place it shows one very important way in which we reach conclusions about what a man knows or believes - i.e. by listening to what he says, reporting what he says in the *oratio obliqua* form, and then provided we are satisfied that he means what he says and is not trying to deceive his audience, moving from 'A said that p' to 'A knows' or 'believes that p'. Moving on to the second point we may note that we are only justified in moving from 'A said that p' to 'A knows' or 'believes that p' if we are satisfied that A is not only disposed to assert p, but is also disposed to act on p or, to put it another way to take p as a premise in his practical reasoning. But given that this assumption is justified, as it usually is, and provided we know or can work out what are for him the practical implications of the proposition in question, we now have a technique for predicting what a man will do from what he says.

The aspect of a man's behavioural dispositions which can be predicted from the various statements he makes about the situation in which he finds himself is the way in which his behaviour is adapted and modulated to fit the various contingencies and relationships within the environmental situation. What cannot be predicted from such statements is the direction in which the individual's behaviour is likely to move in

relation to these environmental contingencies. Thus as we have already seen, we cannot predict what will be the practical implications of accepting a given proposition as a basis for action, unless we know what is acceptable and unacceptable to the individual as a consequence of his actions, in other words what it is that he wants to bring about or prevent. In order to assess the practical implications for his own behaviour of the various propositions which a prospective agent asserts or is inclined to assert about the environmental situation confronting him, we need to select from amongst these various propositions those which mention or have implications for the coming or bringing about of those environmental contingencies which are acceptable or unacceptable to the agent. This means that we have as it were, to look inside the propositions which a man is said to believe and pick out those which contain concepts under which those things which the agent finds acceptable and unacceptable fall or descriptions to which they answer. This I take it, is why the verbs of volition take as their objects neither an actually existing entity or state of affairs, nor a proposition, nor the concept or description itself, but something or anything which falls under or answers to the concept or description used in specifying the object of desire, repugnance or intention, whether or not any such thing actually exists or is capable of existing.

Mentalism as a scientific theory

The way in which propositional attitudes (knowledge and beliefs) combine with wants to yield prediction as how an individual is likely to behave which cannot be derived from a knowledge of either in isolation gives to mentalist explanations of behaviour some of the character and logical complexity of a scientific theory; so that one might be tempted to describe such explanations as forming a naturally occurring psychological theory. It may be said to resemble the more familiar forms of scientific theory in the following respects:

(1) It is designed to enable the user to predict and control the occurrence of events of a certain kind, in this case the particular actions of human beings.

(2) It follows the hypothetico-deductive principle, in that it starts from assumptions about the knowledge, belief and desires of the agent and deduces his probable line of action therefrom.

(3) The initial hypotheses about the individual's knowledge, beliefs and desires from which specific events (his action) are deduced have the characteristic of generality in that many different predictions can be deduced from a given hypothesis about an individual's knowledge, beliefs or desires, when combined with different hypotheses of the same general kind, as in the case instanced by Geach (2) where the behaviour to be predicted in a man who believes that it is going to rain will be quite different according to whether his objectives are those of a gardener, a man wanting to walk from A to B without getting wet or, as in Dr Johnson's case, the desire to do penance.

(4) Despite popular philosophical belief to the contrary, mentalist explanations resemble scientific theories in describing the relations between some, at least, of the causal factors which bring about the outcome predicted. What a man knows, believes, wants or intends can quite properly be described as a cause of his behaving as he does in accordance with the principles which we discussed in Lecture 5, if and in so far as it is the case that he would not have acted as he did if he had not known, believed, wanted or intended to do what in fact he knew, believed, wanted or intended to do.

As against this there are certain respects in which mentalist explanations undoubtedly differ from the kinds of scientific theory to which we are accustomed in the natural sciences:

(1) Although mentalist explanations are designed so as to permit the prediction and control of human action, the means by which this is achieved is very different from that of most scientific explanations and theories. In scientific theories prediction and control is achieved through knowledge of the relationship between the independent variables which determine the events to be predicted and controlled and the dependent variables in terms of which those events are measured. In a mentalist explanation the dependent variable, the action of the individual, is predicted, not so much from a knowledge of the independent variables controlling that event, as from a knowledge of another dependent variable, namely what the agent has to say about the situation confronting him or what he may suppose that he would say if he were asked to give his opinion. The means of controlling human action which is made possible by such explanation and prediction is likewise indirect. By

drawing attention to those aspects of an individual's verbal behaviour which predict how he is likely to behave in the situation in which he finds himself it puts us in a position to control or if you prefer to influence what he does by altering where possible, the kinds of things he is inclined to say about the situation through persuasion and argument.

(2) Although this kind of explanation and prediction follows the hypothetico-deductive pattern, the most important part of the argument hinges on conclusions which follow, not from premisses or axioms which state putative relationships between the individual's behaviour and the independent variables which determine it, as from premisses concerning the environmental situation confronting him which are or would be asserted by the agent himself. In other words we predict the agents behaviour by as it were, reconstructing the inferences which he draws in reaching a decision as to how to act in the situation in which he finds himself on the assumption that, if a conclusion follows logically from premisses which he accepts, a rational man is likely to draw that conclusion.

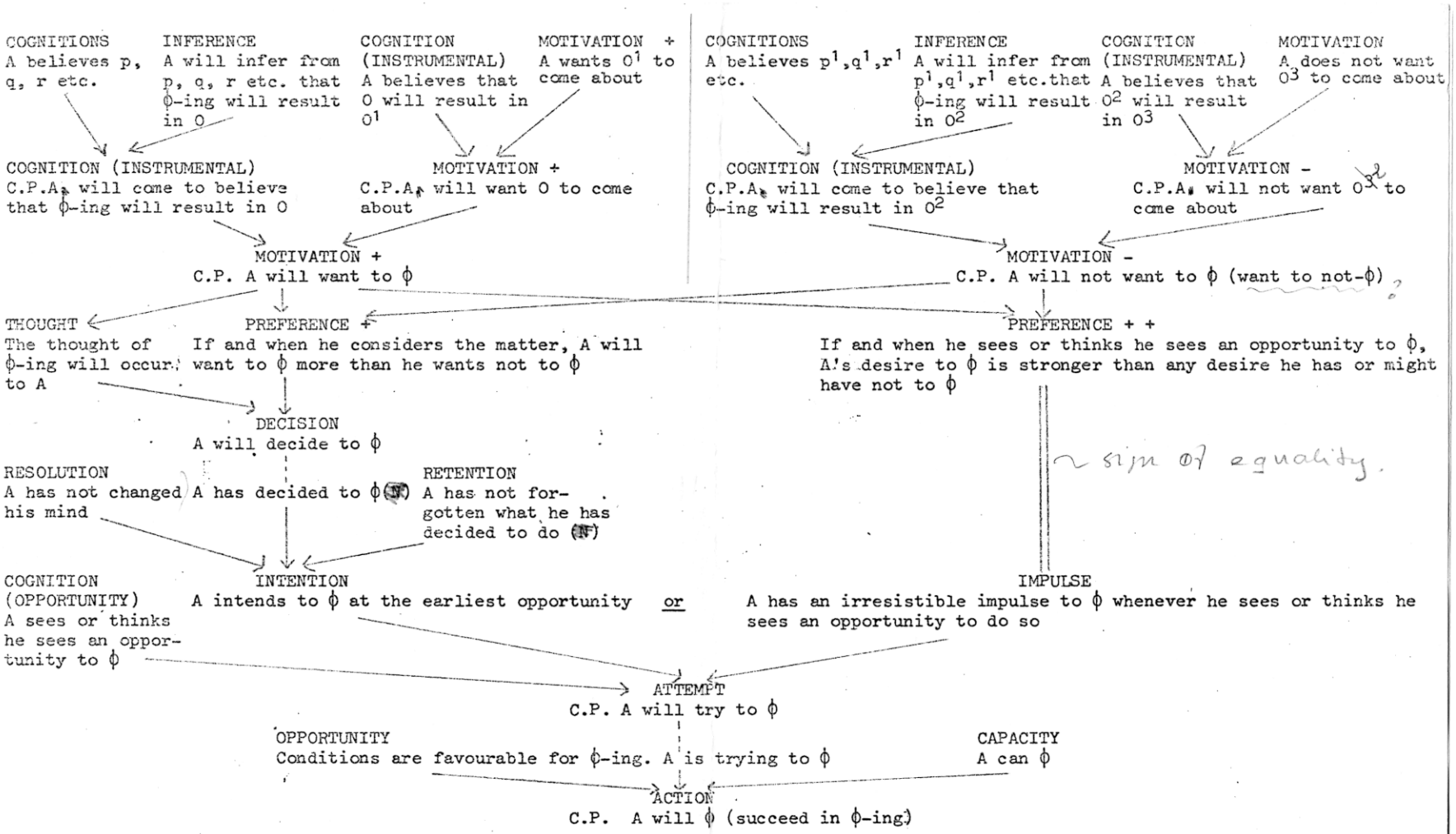
(3) Although the hypotheses about an individual's beliefs and wants from which his behaviour is inferred have the characteristic of generality in that different predictions can be drawn from different combinations of the same set of hypotheses, the hypotheses only have application to the behaviour of the individual concerned. Each individual has his own separate and unique set of beliefs and wants.

(4) Although beliefs and wants can in my view, be quite properly regarded as causes or causal factors which determine how an individual will behave, they are not *the* cause of a man's behaving in the sense of the triggering event which completes the set of conditions sufficient for the occurrence of the action in question. They cannot be said to prompt or precipitate a man into acting as he does, because only an event can do that; and beliefs and wants are not events.

(5) Wants and intentions, together with those beliefs which are most immediately relevant to the explanation of action all involve the agent's, in some sense foreseeing, envisaging or preconceiving, if not the actual effect of his intervention, then at least some other effect which could reasonably be expected to result from such activity or inactivity. In recent years we have become familiar with human artifacts, like guided missiles whose working can be explained and have indeed been constructed on the basis of the laws of physics but which nevertheless share this characteristic of acting in accordance with a preconceived plan, purpose of or intention. It remains true nevertheless, that mentalist explanation is the only system of explanatory concepts which incorporates this feature into the structure of explanation itself.

(6) Mentalist explanations of human behaviour are also unique in that the entities whose behaviour they explain (human beings) make almost as much use of such explanations in giving an account of their own behaviour as they do in explaining and predicting the behaviour of other people. The individual himself is often though not invariably, in a better position to determine what he wants and what he believes, than other people are. On the other hand, although in deciding what he is going to do, he often goes through a process of reasoning similar to that which another person would try to reconstruct in attempting to predict his behaviour from the outside, as it were, he does not usually make use of these explanations in order to understand himself and predict his own behaviour in the way an outsider needs to do. When he gives a mentalist type explanation of his own behaviour, he is usually concerned to justify his own rationality or the morality of his motives and intentions in the eyes of other people. Consequently the *reasons* that he gives are not always the *real reasons* why he acts as he does.

Reasons it has often been alleged by philosophers (4) are not causes. The only point of substance behind this observation in my view, is that when a man gives his reasons for acting as he does, he merely asserts the propositions he believes and these propositions, together with whatever it is that he wants, constitute his reasons for so acting. Now a proposition, as we saw in Lecture 2, is a kind of utterance. It is therefore a universal and not a particular; and as we saw in Lecture 3, kinds or universals can only be said to exist in so far as their instances exist, occur or are as a matter of fact the case. But if propositions cannot properly be said to exist, occur or be as a matter of fact the case, they cannot be a cause of the existence or occurrence of anything either. The asserting or believing of a proposition, on the other hand *is* something that can be


COGNITIONS
A believes p, q, r etc.
INFERENCE
A will infer from p, q, r etc. that ϕ-ing will result in O
COGNITION (INSTRUMENTAL)
A believes that O will result in O^1
MOTIVATION +
A wants O^1 to come about
COGNITION (INSTRUMENTAL)
C.P.A. will come to believe that ϕ-ing will result in O
MOTIVATION +
C.P.A. will want O to come about
MOTIVATION +
C.P. A will want to ϕ
COGNITIONS
A believes p^1, q^1, r^1 etc.
INFERENCE
A will infer from p^1, q^1, r^1 etc. that ϕ-ing will result in O^2
COGNITION (INSTRUMENTAL)
A believes that O^2 will result in O^3
MOTIVATION
A does not want O^3 to come about
COGNITION (INSTRUMENTAL)
C.P.A. will come to believe that ϕ-ing will result in O^2
MOTIVATION -
C.P.A. will not want O^2 to come about
MOTIVATION -
C.P. A will not want to ϕ (want to not-ϕ)
THOUGHT
The thought of ϕ-ing will occur to A
PREFERENCE +
If and when he considers the matter, A will want to ϕ more than he wants not to ϕ
PREFERENCE + +
If and when he sees or thinks he sees an opportunity to ϕ, A's desire to ϕ is stronger than any desire he has or might have not to ϕ
DECISION
A will decide to ϕ
~ sign of equality.
RESOLUTION
A has not changed his mind
A has decided to ϕ (N)
RETENTION
A has not forgotten what he has decided to do (N)
COGNITION (OPPORTUNITY)
A sees or thinks he sees an opportunity to ϕ
INTENTION
A intends to ϕ at the earliest opportunity
or
IMPULSE
A has an irresistible impulse to ϕ whenever he sees or thinks he sees an opportunity to do so
ATTEMPT
C.P. A will try to ϕ
OPPORTUNITY
Conditions are favourable for ϕ-ing. A is trying to ϕ
CAPACITY
A can ϕ
ACTION
C.P. A will ϕ (succeed in ϕ-ing)
C.P. = Ceteris Paribus

said to occur or exist and is therefore, something which *can* quite properly be said to cause other states or events. No such distinction between the reasons for and the causes of behaviour can be drawn in the case of the things that a man wants, since it is not what a man wants that is his reason for acting as he does, but the fact that he wants it.

The Logical Structure of Mentalist Explanations

I have attempted to set out the formal logical structure of Mentalist Explanations on the enclosed diagram. The structure, as I have laid it out, takes the form of a tree with the conclusion of the argument - a sentence frame of the form 'A will Φ' - at its base and with sentence frames representing the various premisses or axioms from which a proposition of that form is directly or indirectly deducible branching off from the main trunk or from one of the branches at different levels.

In explaining a piece of behaviour that has already occurred we move upwards from the base to explanations whose generality increases the greater the number of steps separating it from the base. In predicting behaviour we deduce its occurrence from premisses which may be drawn from a variety of different levels in the hierarchy depending on the available information. In principle the argument can be extended indefinitely in the upwards direction by specifying more general objectives which the agent wants to achieve or more general contingencies which he wants to avoid with respect to which his more specific objectives are interpreted as *means* in relation to the general goal as *end* in the light of the instrumental beliefs which can be attributed to him.

The methodology employed in constructing this tree is to begin at the base and try to find the minimal explanation required in order to account for the occurrence or state of affairs in question i.e.: we look for a proposition which although it does not by itself entail the existence or occurrence of the action or mental state in question, can nevertheless be made to entail this conclusion with the smallest number of auxiliary assumptions. We then write down this explanatory proposition together with the auxiliary assumption or assumptions required to deduce its existence or occurrence. Each of these premisses is then examined with a view to discovering the minimal explanation that will account for the existence or occurrence of the states of affairs adverted to in this explanation of the base level occurrence. The procedure is then repeated at the next level above.

What this procedure is designed to yield at each stage is a set of conditions which are jointly sufficient, in a causal sense of 'sufficient', for the existence of the occurrence or state of affairs specified in the propositions at the level below. In many cases we find that there is more than one set of conditions sufficient for the existence or occurrence of the behaviour mental act or mental state in question. For example if we ask why someone believes a given proposition there are at least four different kinds of possible explanations of why he should have come to hold such a belief: (a) he may have interpreted something that he saw, heard, smelt, tasted or felt in this way (b) he may have been told that the proposition in question was true by someone else and accepted what he was told (c) he may have inferred the proposition from other propositions he believes (d) he may simply have believed what he wanted to believe. Of these various possibilities I have only included in the tree the case where his belief comes about as a result of an inference drawn from other pre-existing beliefs, since it is only in this case that it is possible to state the sufficient conditions for the existence of a belief in terms of our common sense understanding of the matter with sufficient precision to allow the deduction of a conclusion of the form 'A believes p'. This has not been filled in in the case of the belief that an opportunity to Φ has arrived, since such beliefs are not usually arrived at in this way. In the case of the conclusion 'A will try to Φ', two sets of sufficient conditions are shown on the diagram, one for the case where the agent is acting on a preformed intention to Φ and another for the case where he is acting on an irresistible impulse to Φ without such a preformed intention. In both cases the conclusion 'A will try to Φ' is deduced from either of these two premises with the same auxiliary assumption to the effect that 'A sees or thinks he sees an opportunity to Φ'.

In other cases for example, in the case of the deduction of the conclusion that A intends to Φ from the premisses 'A has decided to Φ', 'has not changed his mind' and 'has not forgotten what he decided to do', the causal conditions are necessary and not merely sufficient since, as we have seen in the previous lecture, a disposition to Φ is describable as an intention to Φ only in so far as it has come about as a result of a decision to Φ.

It will be noted that in those cases where what is predicted is an event whose coming about may be arrested by some unpredictable contingency intervening between its initiation and its completion, the prediction is always subject to a *ceteris paribus* (c.p.) or 'other things being equal' clause, which allows for the failure of the prediction in such cases. Given this qualification however, all the conclusions indicated [in] the diagram, with one notable exception, follow strictly according to the canons of deductive inference from the premisses I have indicated. The one glaring exception [is] the case of the arguments at level 5 where no conclusions about the relative strength of the desire to Φ and the desire not to Φ can be inferred from premisses which do no more than predict that the individual in question will have both of these incompatible desires.

The inability to predict the absolute or relative strengths of the desire to Φ and the desire not to Φ is perhaps the most striking defect of common sense mentalist explanations of behaviour considered as a method of generating precise and unambiguous predictions of what a man will do under given circumstances. It cannot be too strongly emphasised however, that this defect is due to the lack of any technique available at the level of common sense for quantifying and measuring the absolute or relative strength of a man's desires and the degree of confidence with which he holds his various beliefs and draws the inferences that he does draw from them, and not to any fundamental incompatibility between concepts like 'believing' 'wanting' and 'inferring' on the one hand and the principles of measurement and quantification on the other. There is therefore no reason in principle why the psychologist should not set out to supplement and improve the mentalist explanation of ordinary language by supplying techniques for quantifying and measuring such things as the subjective probability of a given outcome and the relative, if not the absolute, utility value in terms of gain or loss of different outcomes for a given individual, as is done in statistical decision theory (1).

References

1. W. Edwards - The theory of decision making - *Psychological Bulletin* 1954, 51, 380-417
2. P.T. Geach - *Mental Acts* London, Routledge/Kegan Paul 1957, Chap. 4
3. P.T. Geach - *Logic Matters* Oxford, Blackwell 1972, pp. 139 ff.
4. R.S. Peters - *The Concept of Motivation,* London, Routledge & Kegan Paul 1958
5. G.Ryle - *The Concept of Mind* London, Hutchinson 1949.