PHILOSOPHICAL LECTURE

PREDICTING AND DECIDING

By DAVID PEARS

Read 17 June 1964

OTHER people's decisions can be predicted inductively. But can anyone treat his own decisions in this way? It has been claimed[1] that the answer to this question would be a step towards the solution of the problem of free will. But my aim is at something closer. I ask the question because it opens a way to a problem about the nature of deliberation. How is one person's deliberation related to another person's prediction of its result?

If someone tries to make an inductive prediction of the result of his own deliberation, it looks as if he is trying to see the matter as another person would see it. But can he really take a spectator's seat? Certainly he can, when what he predicts is that, even if, after deliberation, he decided on an action, and never changed his mind, he still would not perform it. But that is not the point. The point that some philosophers[2] want to make is that, when he thinks that his action will depend on his decision, he cannot predict it inductively, because he cannot predict his own decision inductively. They maintain that, in such cases, what he will do must remain an open question for him until he has made his decision. It would follow that an inductive prediction, made by him, of his own decision could never come true, since his prior certainty would exclude the possibility of his subsequent decision. It is admitted that the decision would be possible if he forgot, or ceased to believe, his inductive prediction after he had made it; and that, when he seems to be making an inductive prediction of his own future decision, he may really be making a present decision, expressed in a misleading way. But, it is contended, a prediction of a decision which is genuinely inductive, and made, remembered, and still

[1] By Professor Hampshire in his book *Thought and Action*, chaps. ii and iii.

[2] e.g. Professor Hampshire loc. cit.: Professor Hart and Professor Hampshire in their article 'Decision, Intention and Certainty' in *Mind* 1958: and D. M. Mackay in his article 'On the Logical Indeterminacy of a Free Choice' in *Mind* 1960.

believed by the agent, is strictly self-frustrating. He may take a spectator's seat in such cases, but, so long as he stays in it, he cannot play the whole of his predicted part as agent.

There are two things which make it difficult to assess this answer to my question. First, deciding and acting may be almost simultaneous, and, even when they are not, deciding need not be a definite event. It is no accident that the present tense of the verb 'to decide' leads a very marginal life outside subordinate clauses. Secondly, it is not always clear when the agent's prediction of his own decision is properly called 'inductive'.

In order to circumvent the first difficulty, I shall begin by considering cases in which the agent would naturally and easily make his decision at a definite moment which precedes the moment of action. What I shall say about these cases can be generalized, without much modification, to cover similar cases where, although the decision does not occur at a definite moment, there would be a time before the action at which he could say that he had decided. Later, I shall say something about the very different situation where he finds it hard to make up his mind before the moment of action. In that kind of situation the best way to secure examples where a decision, or something like a decision, might naturally be expected before the moment of action is to assume that there is some special consideration which makes this necessary. For instance, there might be other decisions which he could not postpone, and which depended on his decision in this matter: or other people might require him to make up his mind, perhaps for a similar reason. But, as I said, I shall begin with simpler cases.

In order to circumvent the second difficulty, I shall confine the initial scope of my inquiry even further. My first cases will all be ones in which the agent's prediction of his own decision will be obviously inductive. I shall leave the more dubious cases, where we should hesitate to call it 'inductive', until later. That will make it possible to isolate one problem, the compatibility or incompatibility of deciding with inductively predicting the decision, and to deal with it first.

One way of securing cases where the agent's prediction of his own decision is obviously inductive is to assume that he does not know all the relevant details of the situation in which he will make his decision. Then there might be special circumstances which made it possible for him to predict it inductively in spite of this gap in his knowledge. In a matter of taste, for example, he might predict that in a certain shop he would decide to buy

what a friend of his, with similar tastes, had just bought, even if he did not yet know what the shop offered. Or, to take a more calculative example, he might predict that in a game of chess, confronted by the same simple position as a friend of similar skill, he would decide to make the same move, even if he did not yet know what the position was. In both these examples his prediction would be obviously inductive. Moreover, so long as he remained unaware what his friend's purchase or move specifically were, there is no doubt that, confronted by the situation—shop-counter or chess-board—he could make each of the predicted decisions. For there is not even an appearance of incompatibility between predicting one's own decision under one description and making it under another description, provided that one does not know that the two descriptions are uniquely satisfied by the same decision. But suppose that the agent, before he makes his decision, does find out what article his friend bought, or what move he made, so that he can predict his decision under the description under which it will be made. Then, when he is confronted by the situation, can he still make his decision without giving up his prediction of it? This is the controversial question.

When I ask it, I am, of course, assuming that the agent really does begin by making an inductive prediction, and does not begin by deciding to do whatever his friend does. I am also assuming that he maintains his prediction, neither forgetting it later, nor abandoning it, nor modifying it in any way. However, within these limitations, the question can be generalized a little. The agent might begin by predicting that he would decide to do what his friend advised, or what his friend predicted, on the evidence of his (the agent's) past decisions, that he would decide to do. The only restriction on the descriptions in the agent's original prediction is that they must be descriptions from which, given additional information available *before* the decision is made, it would be possible to deduce the description under which it will be made. So the description in his prediction must not be 'what my friend will imitate', if the friend will imitate whatever he does. Of course, if the agent is going to make his decision under a rationalizing description, the matter becomes more complex. But I shall ignore that complexity, and confine myself to cases where he makes his decision under a description which connects it with the desires from which it issues.

What is the answer to the controversial question? Consider first the more calculative case where the agent asks himself what

move he will make when he is handed the chess-board, and predicts that it will be the same move as his friend's, and then discovers what that move was. Here, provided that the position is simple, a high degree of certainty is often justifiable.[1] For this kind of practical problem is not merely like a theoretical one: it actually contains a theoretical one. Now, whether the agent's problem is only theoretical—what move would lead to the swiftest certain check-mate—, or practical—what move to make—, there are things which remain to be done after he has predicted the result inductively, and which cannot be done before he sees the board. The question is what ought we to call these things. Had he not made a certain prediction of the result, we should say that he solved the theoretical problem, and, if there were also a practical one, that he deliberated and decided. But, since he has made the prediction, we cannot say that he solves the theoretical problem, because solving is discovering the solution by working it out. Nor can we even say that he is checking the solution. For, given his initial inductive certainty, he will be working it out from the position in order to see *how* it fits, rather than *that* it fits. Still, this is something that is related to solving. It is what is left when initial uncertainty is subtracted from solving.

If his problem is practical, what he does will be slightly more complex. For he will start not from the position alone, but from the position and his desire for the swiftest certain check-mate, and he will work forward from these two to the project. However, as before, he will be seeing how the project fits rather than that it fits. But this time that will not be all that he is doing. Something that is not purely intellectual will be happening simultaneously. His desire will be directed on to the project. And there is a great difference between knowing that this will happen and actually feeling it happen. Can we call this deliberating and deciding? Perhaps not. But, if we do not, it is important to see that this time more that is the same is left when the initial uncertainty is taken away. For, though he may not be making the decision, nevertheless, when he sees how the project fits his desire and the position, and when he feels the direction of his desire, he is making the decision his own. No such essential part is played by desire when the problem is theoretical, since, though he may not want to solve a theoretical problem, the solution does not depend on his desires. So, if, in my example, he makes the solution of the theoretical problem

[1] But the degree of justifiable certainty is limited, see p. 222.

his own, the sense in which he makes it his own will not be so strong as the sense in which he makes the decision his own when the problem is practical. Hence the subtraction of initial uncertainty from normal deciding leaves more that is the same. What about deliberating? There too, I think, the same considerations apply. Only, we should add that, since deliberation is a sort of working out, if we refuse to allow that an agent can deliberate with prior certainty, this refusal ought to be even more qualified.

The other example, where a choice between available articles is a matter of taste, is different in several ways. Desires that may vary from person to person play a larger part, and calculation plays a smaller part. Consequently, it would often be artificial to try to extract a theoretical problem from the practical one: exposure to the articles is almost inevitably followed by the process which has a claim to be called deliberating; and prior certainty is more rarely attainable. Still there are cases of this kind where a high degree of prior certainty is attained, and, if we ask whether the agent deliberates and decides in such cases, the answer will be much the same as before; except that, since competing desires play a larger part here, a negative answer would need to be qualified even further.

It has been suggested[1] that the idea, that there is anything like a contradiction concealed in the phrase 'deciding to do *A* with prior inductive certainty that one would decide to do *A*', is an illusion; an illusion which comes from thinking that the agent decides to do *A* in order to achieve certainty that he will in fact do *A*. For, if we think this, we shall naturally regard his inductive certainty that he will decide to do *A* and his actually deciding to do *A* as two competing, and therefore, perhaps, incompatible ways of achieving certainty that he will in fact do *A*. But, it is contended, people decide in order to achieve certainty about what to do, and not in order to achieve certainty about what they will do. And to those who realize this, it is suggested, the phrase will no longer even appear to be contradictory.

If this is correct, my treatment of the two examples is too cautious and qualified. But I do not think that it is correct: not just because people do sometimes decide to do *A* with the primary purpose of achieving certainty that they will in fact do *A*, and then building on it; but for the more important reason

[1] By J. W. Roxbee Cox in his article 'Can I know beforehand what I am going to decide?' in the *Philosophical Review* 1963.

that, whatever the primary purpose of a decision may be, after it has been made, the agent will be certain what he will in fact do, and so the apparent contradiction cannot be removed in this way. In any case another apparent contradiction confronts us when we consider certainty about what to do. For in my two examples, if the agent assumes that his friend made the right choice, he will be inductively certain about what to do: and it is equally plausible to maintain that there is another contradiction concealed in the phrase 'deliberating and deciding with prior inductive certainty about what to do'. But, as I have been arguing, even in my two examples there is room for important elements of deliberating and deciding. However, it is true that in unusual cases, like these, the agent, before he is confronted with the actual situation, will not have the feeling that normally accompanies certainty about what to do. He cannot have it until he makes the decision his own by seeing how it fits the situation and his desires, and by feeling the direction of his desires. These are the elements of deliberating and deciding that come later.

I hope that a fairly general truth is beginning to emerge. To put it negatively, if an action depends on a decision, it is an exaggeration to say that the two things which yield certainty about it, deciding and predicting the decision inductively, are independent and uncombinable. To say that they are independent is exaggerated, because the description under which the decision is predicted must be one from which, given additional information available before the decision is made, it would be possible to deduce the description under which it will be made (unless the description under which it will be made is a rationalization). To say that they are uncombinable is exaggerated, because one of them, unmodified, can be combined with a modified form of the other. Even if the agent maintains the so-called spectator's viewpoint, he will not be prevented from playing his part as agent: he will only play it rather differently (unless, of course, he has a desire to falsify the prediction as such: but I am assuming that he has not).

It might be admitted that the two things that yield certainty about an action that depends on a decision can be combined, after some modification of one of them, in unusual cases where the agent predicts his decision under the description under which it will be made before he discovers the relevant features of the situation of choice. For, when he discovers them, he does something very like deliberating and deciding, and, even if he does

not make his prediction come absolutely true, at least he catches up with it. But in the more usual cases, where he already knows the relevant features of the situation of choice, it is not so easy to see how his prediction can outstrip his deliberating and deciding and yet come almost true. However, this often seems to happen. How can it happen? This time there is an additional difficulty. For how can his prediction keep its inductive character, in spite of the fact that it goes through his own desires and knowledge of the situation of choice? How can it avoid becoming a decision made in advance.

I think that it is clear that we cannot go very far towards answering this question without examining the deeper operations of induction, and unearthing contingent facts which are taken for granted in everyday life and built into the structure of our concepts. However, I shall begin, as before, by taking cases in which the agent's prediction is obviously inductive, and I shall assume that naturally, and without the pressure of any special consideration, he would make, or at least would have made his decision before the moment of action. Now where the deliberation is largely calculative it will be difficult to find such cases. For the more calculative the deliberation, the more unlikely it will be that the agent's prediction will outstrip his decision and still keep its inductive character. For instance, if his desire to check-mate swiftly is firm, and, if he sees the position, and, after calculating, appears to predict with complete certainty that he will decide to make a particular move, there are strong reasons for saying that he has already decided, but is expressing his decision misleadingly. For it is irrelevant that he has not yet made the move or touched the piece, and any behavioural confirmation of the hypothesis that he had not yet made the decision would be exceedingly likely to undermine the hypothesis that he was certain that he would make it. So I shall choose examples where the deliberation is very far from being purely calculative, and involves the weighing of desires that may change.[1] In such cases there will be more than an analogy between predicting one's decision before it is made and predicting one's emotion before it is felt.

In this category there seem to be two types of case where the agent's inductive prediction outstrips his decision. First, there is the radical type of case where he predicts that the general pattern of his desires will change, and that after it has changed

[1] Even the desire to win a game of chess may come and go. But usually its constancy is taken for granted in deliberation about one's next move.

he will make a particular decision. For instance, he predicts that it will change after physical or psychological treatment, conversion, or some other cardinal experience. Secondly, there is the less radical, and far more frequent type of case, where he predicts only that, in some particular matter, his present desire[1] or favour will change, and that he will make a different decision in the end. Everyone would agree that this happens in matters of taste and in cases where pleasure is the avowed aim. But it also happens in other kinds of delicately balanced predicament. If we were interested only in the nature of desire, the difference between the two types of case, the more and the less radical, could be presented as simply a difference of degree. For even a desire about a particular matter contributes something to the general pattern, and a change in it might be part of a larger upheaval. But, since we are also interested in the agent's ability to predict the change inductively, the difference is, perhaps, more than one of degree. For he can often predict from his own past record that his desire in a particular matter will change, but the prediction of a change in the general pattern would need to be based on a striking external cause.

What the two types of case have in common is that the agent predicts a change in his desires. He considers the possible projects, favours one of them most, and then predicts that he will decide on a different one. In the radical type of case the prediction, which is based on an external cause, is obviously inductive, and it will sometimes yield a high degree of prior certainty. If it does, will the agent be able to make the prediction come true by deliberating and deciding? I think that he will, but not quite in the way that he could in my first two examples. For in those examples, when he made the predictions, he did not know the situations: but in this case, when he made the prediction, he already knew the situation and saw how the project would fit it and the pattern of his desires, if that pattern changed. What happens later is that it does change, and the direction of his desires, which he then feels, is new. In general, in the triangle formed by situation, desire, and project, either the first point is not known by the agent when he makes his prediction, or the second point is not fixed.

The usual objection to this answer is that the agent decides, or ought to decide, before the change comes about. But ought he? Surely the idea that beneath such changes there is an

[1] I use the word 'desire' in an inclusive way. Contrast the exclusive use of the verb 'to want' in 'deciding to do what one does not want to do'.

unchanging source of decisions is a moralizing fiction. How can he make a decision in advance? Perhaps it will be suggested that he can decide in advance to do whatever he feels like doing later. But, when a decision is expressed in that way, there is an implied contrast between one's own later feelings and other considerations, and the decision, which issues from the present pattern of desires, is a decision to exclude those other considerations. Our case, however, is quite different. In it the implied contrast is lacking, and so that way of expressing a decision in advance would be deprived of its usual point. Moreover, the decision could only issue from a higher desire, which could not compete with the others, the desire to be true to oneself.

In the less radical cases, where the agent claims that in a particular matter his desire will change, it might be doubted whether the claim is really inductive. As before, he knows what the possible projects are, favours one most at the moment, but predicts that in the end he will decide on a different one. But this time there is no suggestion that the general pattern of his desires will change, and the prediction is based on the outcome of his own previous deliberations in similar situations. So his reasoning is very closely connected with the usual operation of the pattern of his desires. However, it can still be called inductive. If he had waited for his desires to point in their final direction, his reasoning would not have been inductive. But he does not wait. His prediction outstrips his decision, and so, though it is closely connected with the usual operation of the pattern of his desires, the connexion is not the kind of connexion that would deprive his reasoning of its inductive character.

Can he, in this kind of case, make his prediction come true by deliberating and deciding? The answer seems to be that, even if he is quite certain of his prediction, he can, in much the same way that he could when the change was more radical. However, in this kind of case he would seldom in fact feel very certain about his prediction. So I ought also to ask my question about cases where his prediction is more tentative.

Tentative predictions introduce complications which are off my route. So I shall deal with their effect briefly and schematically. When the agent predicts confidently that he will decide to do A, let us suppose that he would assign the probability $1/x$ to the proposition that he will decide to do A. When he predicts it tentatively, the probability that he would assign to it would be smaller, say $1/(x+w)$. Now nobody would claim that it is absolutely certain that, if he decided to do A, he would do it, even if

the moment of action were very close. For, even if nothing else changed in the interval, his desire might change. Of course, there are cases where any change would be enormously improbable, particularly if the interval were short. Let us say that in such cases the agent's decision to do A would give the proposition that he will do A the probability $1/z$. I pointed out earlier that often deciding and acting will be almost simultaneous. When this is so, $1/z$ will be almost indistinguishable from 1.

There is another complication which ought to be mentioned at this point. There is something else, which is very like a decision, but less firm. It may be that decisions are, by definition, firm. If so, the other thing ought to be called 'a tentative intention'. The noun 'intention', and the verb 'to intend', even when they are not qualified, often have this suggestion of tentativeness. Now suppose that an agent predicts inductively that he will form a tentative intention to do A.[1] Then the formation of this intention would give the proposition that he will do A a probability less than $1/z$, say $1/(z+y)$. These two assessments of probability can be made by anybody, but I am assuming that they are made by the agent.

Now, if we hear him predict that he will decide to do A, or that he will form the tentative intention of doing A, we may inquire what probability he would assign to the proposition that he will in fact do A. So far, I have only taken cases where the prediction is confident and what is predicted is a decision, and in such cases he would assign the high probability $1/xz$. But there are also three other theoretically possible types of case. He can confidently predict the formation of a tentative intention, and then the probability that he would assign would be $1/x(z+y)$: or tentatively predict a decision, and assign the probability

[1] We might ask when, according to him, this tentative intention would be formed. In the cases so far examined it was natural to expect that he would have decided before the moment of action arrived. But suppose that he is going to find it difficult to make up his mind. Then why should he predict that there will be some moment before the moment of action at which he will have formed a tentative intention? Would it be a final tentative intention? As the moment approached closer to the moment of action it would become increasingly absurd for him to predict that the intention that he would then form would be only tentative. So how could he ever predict that he would form a tentative intention, unless he were using this phrase only as a synonym for the gradual emergence of a preference?

I think that the answer to this question is that, though he might be using the phrase in this way, he need not be. For there might be some special consideration which would force him to crystallize his desires, however inadequately, before the moment of action arrived. Cf. p. 194.

$1/z(x+w)$: or tentatively predict the formation of a tentative intention, and assign the probability $1/(x+w)(z+y)$.

This fourfold schema generalizes the problem, and my original question can now be put in its most general form: can the agent make a decision or form an intention which, in his estimation, gives the proposition that he will perform the action a probability scarcely greater than the probability which he had already implicitly assigned to it when he inductively predicted the decision or the formation of the intention? I have been arguing for a qualified affirmative answer in some cases where the prediction is confident and what is predicted is a decision. Exactly the same arguments apply when the prediction is confident and what is predicted is the formation of a tentative intention. But in the other two types of case, where the prediction is tentative, the situation is quite different. For here, when the predicted decision is actually made, or the predicted tentative intention is actually formed, it will give the proposition that the agent will perform the action a probability substantially greater than the probability which he had already implicitly assigned to it when he made the inductive prediction. Consequently, in these cases, the agent's prior certainty is not great enough to modify the nature of his deliberation and decision, or the formation of his tentative intention. This explains why people find nothing puzzling in the very frequent cases where an agent predicts with less than complete certainty that in a particular matter his desire will change, and that he will make a different decision or form a different intention in the end.

The existence of these cases, which is hardly in dispute, is enough to dispel two very common prejudices. The first is the idea that an agent cannot really make an inductive prediction of his own future decision, since, if he did, that could only be because, between the moment of prediction and the moment of decision, there was going to be such a change in him that the decision would not be, in the full sense, his own. The second is the idea that, if he considers the possible projects and favours one most, he cannot help deciding on it. The first is connected with the moralizing fiction that I mentioned just now.[1] Against the second it is enough to point out that, even when the result of deliberation is a very strong preference for a particular project, the preference need not amount to a decision. Everyone would agree that it need not in cases where the

[1] It is the other horn of the dilemma: 'Either there is an unchanging source of decisions, or they will not really belong to the same person.'

agent thinks that the circumstances may change, or that the matter may be taken out of his hands. What is less obvious is the thesis for which I have been arguing, that, even if there are no considerations of this sort in his mind, his preference still will not amount to a decision if he has inductive reasons for thinking that it will change.

I asked my original question in the hope that the answer to it might throw some light on the way in which one person's deliberation is related to another person's prediction of its outcome. My idea was that, if we could see how the two things fit together in cases where it is the agent who predicts his own decision inductively, that might help us to see how they fit together in other cases. In all cases alike the difficult thing is to see how inductive and immediate knowledge are related. Now two distinct questions are possible here. First there is a question about deliberation. I pointed out earlier that, if what is predicted inductively is the agent's decision, then, provided that it is not made under a rationalizing description, its description in the prediction must be one from which, given additional information available before the decision is made, it would be possible to deduce the description under which it will be made. Obviously this also applies to an inductive prediction that the agent will favour a particular project most. But then what is the relationship between the agent's immediate knowledge, based on deliberation, of his own selective favour and perhaps decision, and the prediction that he would favour that project most and perhaps decide on it, which was based on inductive reasoning? That is the first question, which is about the process of deliberation. The second question concerns what follows deliberation. What is the relationship between an inductive assessment of the probability that the agent will carry out a particular project, given that he has just decided to do so, and his own immediate knowledge that he will carry it out?

I shall answer these two questions briefly and dogmatically. My answer to the first is that, if the dispositional theory of desire allowed that a desire might be manifested in the inner life of the agent as well as in his behaviour, it would explain the relationship between the two kinds of knowledge and reasoning. My answer to the second is that the concepts of intention and decision are founded on very general contingent connexions, and that, when their substructure is analysed, the relationship between the two kinds of knowledge becomes intelligible. It would take more than my remaining time to defend these two

answers adequately. So I shall merely sketch a defence, which, I think, might become powerful if it were elaborated.

I take the second question first. I have argued that the agent can predict inductively that his desires will change, and that he will decide on a different project from the one that he favours most at the moment. But then it looks as if, in the straightforward cases where he simply decides to do what he favours most at the moment, he must be making the inductive assumption that they will not change. Otherwise how could he be so certain that he will do it? Now there are various things which might make him uncertain. There is, for instance, the possibility that he might be prevented. But, of course, I am not concerned with that kind of possibility, but only with the possibility that his desires might change. And even that is too wide, since a change in his desires might be produced by a change in the situation, and the possibility which really exposes my difficulty is the possibility that his desires alone might change. My point is that it looks as if he must be making the inductive assumption that at least this will not happen, and naturally this assumption would sometimes be mistaken.

It will be objected that, if he takes no account of things outside himself, his certainty that he will perform the action is immediate. If this means that normally he will not use the kind of inductive argument that I have been describing, it is true. But that might show only that he assumes that his desires will not change. Admittedly, if others build on his decision, or if he himself does, a new desire will be brought in, the desire not to disappoint them,[1] or the desire not to upset his own further plans, and each of these will help to keep his favour constant. But perhaps the introduction of these new desires would only complicate his assumption without altering its inductive character.

Decisions should not be idolized. They are not very different from desires, and their effect on the future and the foreknowledge which accompanies them are not magical. There is often only a very slight difference between favouring a particular project most and deciding on it. It may be only that one dismisses the matter, perhaps merely because there is no more time to consider it. And, if we are impressed by the connexion between a decision and the future, at least we ought not to be

[1] There is a difference between keeping up with a statement of intention and keeping a statement of intention close to what one would have done even if one had not made it, and there is room for two distinct virtues here.

superstitious about it. The announcement of a decision may be a sort of certificate, but the decision itself is not one. There really does not seem to be any reason to reject the view that even in the straightforward case, where an agent claims to know that he will perform a particular action, his claim, although it is immediate, is founded on a piece of inductive self-knowledge. Of course he may not be exceptional, and what he claims to know about himself may be merely that his constancy is average. But it could be something more interesting than this, since people vary, and it could be mistaken.

There are two theories which deny, or seem to deny this contention, or at least part of it. According to one, when the agent says 'I will do *A*', this either is a command, and not a statement, or at least it is more like a command than a statement. The other allows that it is a statement, but implies that its eventual truth or falsity,[1] in so far as it depends on things inside the agent, is always ascertainable by him at the moment of utterance. An examination of these two theories might assist the defence of my contention.

Is 'I will do *A*' like a command? Or rather, let us open the bidding at the top, and ask whether it actually is a command. This question has recently been answered in the affirmative.[2] It is said that an expression of intention, like 'I will do *A*', may be regarded as a kind of command addressed to oneself, and that the utterance 'I intend to do *A*', when it is a genuine report of a state of mind, is tantamount to the statement 'I have said in my heart "Let me do *A*"'. The kind of command that is meant must be self-exhortation, which, according to this theory, in the latter case, is said by the agent to have been done by himself in the past, and, in the former case, is actually being done by him audibly at the moment. But how can the theory allow for the fact that he might be insincere in what he says? When he says what he has done, he may, of course, be lying. But that is not possible in the other case, in which he does not make a statement at all. Nor does the possibility of a lie completely cover the possibilities of insincerity when he reports his past self-exhortation. For his past self-exhortation may itself have been 'insincere'.

[1] See p. 209 for an explanation of this phrase.

[2] By A. Kenny, in his book *Action, Emotion and Will*, pp. 216–27. However, his thesis, that an intention is a species of command, may be only an emphatic way of saying that the two things are similar to one another. It ought to mean that intentions possess the generic properties of commands and certain specific properties of their own.

The solution proposed[1] is that an insincere expression of intention is a piece of overheard self-exhortation which the speaker does not mean: just as an ordinary 'insincere' command is an exhortation to another person which he does not mean; whereas an insincere statement is one which he does not believe. Similarly, the past self-exhortation, even when he reports it truthfully later, may not have been meant by him at the time.

There are many obscurities in this theory, but the points that I shall make against it are simple. To exhort oneself to do something is a way of getting oneself to decide to do it, or else a way of keeping oneself up to the mark after one has decided to do it: to form an intention to do something is neither of these things. If someone exhorts himself to do *A* in order to get himself to decide to do it, he has not yet fully formed the intention to do it. Consequently, in this case, though it is true that, if he does not really mean his self-exhortation, and if he knows that he is overheard by another person, then that would be a devious kind of insincerity, nevertheless he would not be deceiving the other about his intention to do *A*, but only about his intention to get himself to decide to do *A*. If, on the other hand, he exhorts himself in order to keep himself up to the mark after deciding to do *A*, this piece of self-exhortation comes too late to express the intention to do *A* at the moment when it is formed. Consequently, in this case, though it is true that, if he believed that the self-exhortation was necessary, and if he did not mean it and knew that it was overheard, then that too would be a devious kind of insincerity, nevertheless he would be deceiving the other not about the present formation of an intention to do *A*, but, rather, about the efficacy[2] of that intention, which had been formed in the past: for he would be implying that it needed reinforcing, and yet he would only be pretending to reinforce it. Therefore, when 'I will do *A*' expresses an intention that is formed at the moment of utterance, it cannot be right to regard it as a piece of self-exhortation. It follows that the other half of the theory, which analyses the formation of an intention in the past in a similar way, cannot be right either.

Moreover, even if the thing about which the speaker might deceive his audience when he says 'I will do *A*' or 'I intend to do *A*' had been the thing about which he might deceive them when he produces what the theory regards as the equivalent of these utterances, it is also important that the method of deception suggested by the theory is too devious. For conveying information

[1] A. Kenny, loc. cit. [2] Efficacy of intention is defined on p. 212.

is not the primary purpose of self-exhortation, whereas it is the primary purpose of the two utterances.

So it looks as if we ought not to expect more than an analogy between intentions and commands. The most important point of analogy that has been suggested concerns their direction of fit. It has been said that, when an action does not fit what the agent said that he would do, it is the action that is mistaken and not what he said:[1] and this direction of fit is characteristic of commands, whereas the opposite direction of fit is characteristic of statements. Let us signalize this by calling commands and intentions 'dominant partners', and statements 'subordinate partners'. Then another point of similarity that has been suggested is that intentions, like commands, produce the subordinate partners that fit them.[2] But already it is not clear exactly what the dominant partner here is supposed to be. Commands can be heard or seen, but in the case of intentions many of the candidates for the position of dominant partner are not perceptible. Is the dominant partner the announcement that one will perform the action, or the decision to perform it, or the knowledge that one will perform it, or, perhaps, the intention itself? A third point of analogy that has been suggested[3] is that, if someone says that he will do something, the contradictory rejoinder would not be that he will not, because he never does such things, but, rather, that he will not, because you are going to stop him: just as, it is said, the contradictory of a command is not the prediction that for some reason the thing will not be done, but, rather, another command, not to do it.

In these suggestions too there are many obscurities. My discussion of them will be aimed at establishing only one thing: that, whatever the exact analogy between intentions and commands, it ceases at the point where my problem begins, since it contributes nothing to an account of the agent's knowledge that he will in fact do *A*:[4] indeed, if it is exaggerated, it actually blocks any account of this knowledge.

First, it is true that an action which does not conform to an

[1] By E. Anscombe in her book *Intention*, pp. 55–57. A. Kenny agrees with this—loc. cit., p. 216.

[2] By E. Anscombe, loc. cit., p. 87. [3] By E. Anscombe, loc. cit., p. 55.

[4] E. Anscombe herself makes this point, loc. cit., p. 55. 'But, returning to the order and the description by the agent of his own intentional action, is there not a point at which the parallelism ceases: namely, just where we begin to speak of knowledge? For we say that the agent's description is a piece of knowledge, but an order is not a piece of knowledge. So though the parallelism is interesting and illuminates the periphery of the problem, it

unchanged intention is often a mistake. But it does not follow that the agent did not make another mistake when he said what he would do. He did. Admittedly, if what he said turned out to be mistaken because he changed his mind later, there would not also be a mistake in his action. But it does not follow that the two kinds of mistake are incompatible; nor, of course, if it did, could it follow from this that there never could be a mistake in what he said. The idea that the two kinds of mistake are incompatible comes from assuming that there can be only one direction of fit here. But why? What fits what? Certainly the action (subordinate) fits the intention (dominant). But also, if the agent says that he will perform the action, his statement (subordinate) fits the action (dominant).

Of course, those who deny that the agent can make a mistake when he says 'I will do *A*' do not think that he is using the future tense to make a future perfect report of his present state of mind. Their idea is that there is a logical connexion between sincere utterances of this kind and subsequent performances. This connexion is, up to a certain point, flexible: for failure to perform the action does not prove insincerity if a suitable explanation is forthcoming. But their contention is—and it is this contention that I am now challenging—that, whether or not an explanation is forthcoming, the fact that the action is not performed cannot show that the agent was mistaken in what he said earlier.

Admittedly, in such a situation we are at least reluctant to call what he said 'false', or even 'not true'. For these predicates, used by themselves, home on to a different target, the agent's implication[1] that he intended to perform the action. There seems to be a very reasonable feeling that the front-line target for truth and falsity is the thing that the agent has the best chance of getting right. So I prefer to say that his statement, that he will perform the action, may come out true, or may possess eventual truth. But, whatever semantic phrases are used, the pejorative one may imply that a mistake has been made. Why should a mistake have to be signalized by the word 'False', or by the phrase 'Not true'?

fails at the centre and leaves that in the darkness that we have found ourselves in.'

The point on which I disagree with her is this: it seems to me that she exaggerates the analogy between intentions and commands, and that the exaggeration blocks any account of the agent's knowledge.

[1] It is difficult to determine the logical character of this implication, and in what follows I make no attempt to do so.

To say 'I will do *A*' is, on any view, to hold up a rather complex target. If someone retorts 'You will not in fact do *A* (although you intend to)', that will hit the target. If he retorts 'You do not intend to do *A*', that too will hit the target. How should we characterize these two impacts? The simplest answer to this question is that the target is a conjunction; and that each retort is the contradictory of one of its members. If that answer were right, the contradictory rejoinder would be the disjunction of the two retorts, and the complete denial would be the conjunction of them. However, there are reasons for regarding this simple answer as too crude. I shall not explore those reasons, or try to refine on the answer. But, if what I have been saying is right, any refinement of it must allow the first retort to be characterized as the imputation of some kind of mistake in what was said. This does not require that the first retort should be the contradictory rejoinder.[1]

But at least, it will be said, intentions, like commands, produce actions. But what exactly produces an action? Certainly not the agent's statement, nor even his knowledge that he will perform it. And, in order to see what a small fraction of truth there is in the idea that the decision produces the action, it is only necessary to reflect how much its efficacy would be reduced if one remembered it without remembering the reasons for it. The efficacy of an intention would be similarly reduced if one remembered only that one had formed it, and not why. If, on the other hand, the thesis, that decisions or intentions produce actions, means that desires produce actions through decisions or intentions, there is much more truth in it, but, correspondingly, less room for the analogy with commands.

So, though the analogy between intentions and commands may well be worth exploring further, it clearly does not account for the agent's knowledge that he will in fact do *A*. Indeed, if the analogy is exaggerated and if the similarity between 'I will do *A*' and 'It will rain' is underestimated, any account of this knowledge will be blocked.

The second theory which I said that I would examine is simply the denial of a corollary of my contention. My contention

[1] The idea that the contradictory rejoinder is 'You will not because I am going to stop you' seems to be produced by the requirement that a contradictory rejoinder must be the same kind of utterance as its target, and in the same person (i.e. in this example the first person singular). But if this exceedingly stiff requirement has to be met as well as the usual requirement for contradictories, how can there be such a thing as the contradictory rejoinder to 'I will do *A*'?

is that, when an agent says that he will do *A*, his knowledge that he will in fact do *A* is based partly on the inductive assumption that his desires will not change. It would follow that the eventual truth or falsity of his statement, in so far as it depends on things inside him, is not always ascertainable by him at the moment of utterance. But the second theory, although it concedes that he makes a statement, maintains that its eventual truth or falsity, in so far as it depends on things inside him, is always ascertainable by him at the moment of utterance.

This theory is connected with a particular kind of analysis of sincerity, and it is that analysis that I am going to challenge. The analysis exaggerates the rigidity of the connexion between sincere statements of the form 'I will do *A*' and subsequent performances. Everyone agrees that there must be some flexibility in this connexion, since, even if such a statement were sincere, external things might prevent it from coming out true. But exactly how flexible is the connexion? The theory implies that its flexibility ceases at the point where we begin to consider things inside the agent. It maintains that, if anything of that kind prevented it from coming out true, that could only be because he had spoken insincerely. For insincerity is avoidable distortion of the truth, and, according to the theory, the eventual truth or falsity of his statement, in so far as it depended on things inside him, was ascertainable by him at the moment of utterance.

But how could this always be so? Sometimes his statement that he will perform the action will be a downright lie. But it may be insincere because his reasons for suspecting that he will not are strong enough to make it an exaggeration. Now, if we confine our attention to his inner life, there are two reasons why he may suspect that he will not perform it. It may be that he does not favour the project enough, or it may be that he does, but has reasons for predicting a change in his favour (if so, his present favour, as I pointed out earlier, would not amount to a decision). If the fact is that he will not do it simply because his present favour is deficient, it is very largely correct to say, as the theory does, that he is aware of the eventual falsehood of his statement at the moment of utterance. If the fact is that he will not do it because his favour will change in a way which he is in a position to predict, it is still often correct to say this, or at least to say that he might suspect that it would not come out true. But if the fact is that he will not do it because his favour will change in a way which he is not in a position to predict, it is never correct to say anything of this kind.

The theory that the eventual truth-value of his statement, in so far as it depends on things inside him, is always ascertainable by him at the moment of utterance is another moralizing prejudice. The idea is that, though his favour at one moment may not be a sufficient basis for a decision because it may change later, nevertheless his psyche is a crystal in which any such change may be seen in the future. But how can it be seen? How could such foresight fail to have an inductive basis, explicit or buried? How, then, could the future always be clear to him?

It is understandable that some recent analyses of sincerity should credit the agent with more foreknowledge than he always possesses at the moment when he says 'I will do *A*'. Let us call an intention which will not fail to be fulfilled because of anything inside the agent 'an efficacious intention'.[1] Then what recent treatments of this problem have stressed is that there must be a logical connexion between sincere announcement and efficacy. But, if the connexion is not rigid, the analysis should be cast in the following form: if the intention is not efficacious, the agent was necessarily insincere unless . . . and here a list of escape-clauses is given. But what sort of escape-clauses should be put on this list? If a photographic film does not react to light, it was necessarily insensitive when it was bought, unless it suffered some specifiable change in the interval. But this is a treacherous analogy. For when the agent's favour is strong[2] but his intention is not efficacious, he can be fairly charged with insincerity, but only if he foresaw the later change in himself. It is so easy to forget that this difference between him and a material object, which appears to, but does not really possess a dispositional property, must produce a difference between the two analyses. If this is forgotten, we shall think that we are testing his sincerity when we are not.

This trap is avoided if the analysis of sincerity allows for possible lack of foreknowledge. But, if it does this, it will lose another point of similarity with the analysis of the defect in the film. For in that case the escape-clauses would all mention things that might have happened between the moment of purchase and the moment of use. But that part of the analysis of sincerity

[1] An efficacious intention need not be fulfilled, since something outside the agent might prevent its fulfilment. There could be an external impediment, or there could be a change in the circumstances which produced a change of mind.

[2] Strength of favour is not the same as efficacy of intention. The latter is defined above: the former is explained, but not defined, on pp. 213 and 214.

which allows for the agent's possible lack of foreknowledge of later changes in himself will not necessarily confine itself to developments in his history between the moment of utterance and the moment for action. For instance, we might know from his earlier history that he must know that his present favour is likely to change. This sort of possibility is forgotten by those who put too much trust in the analogy between the two analyses.

The sincerity of the agent's statement cannot be connected with the efficacy of his intention in the way proposed by some recent analyses. But perhaps there is something else which can be, namely, the strength of his favour at the moment of utterance. Can we produce an analysis of strength of favour which will be closely analogous to the analysis of the sensitivity of the film? Perhaps we may say that, if an intention is not efficacious, then the favour was not strong at the moment of utterance, unless . . . and here will follow a list of escape-clauses, all of which will mention possible developments in the history of the agent between the moment of utterance and the moment for action. For instance, he might definitely change his mind, or suffer from aboulia. If this schema for the analysis is correct, the strength of an agent's favour is very nearly a necessary but far from a sufficient condition of his sincerity when he says 'I will do *A*'. It is very nearly a necessary condition, because his statement is at least exceedingly unlikely to be sincere if his favour is weak:[1] it is far from a sufficient condition, since, even if his favour is strong, he may expect it to change.

But what would happen if his favour seemed to be strong and his intention turned out to be inefficacious, and yet none of the escape-clauses applied? Either we would say that his favour was not really strong, or we should have to admit that the explanation eluded us, and that the list of escape-clauses was not complete. The choice between these two alternatives would be difficult. For, if we believe that it is almost inconceivable that he should make a mistake about the present strength of his favour,[2] his sincerity when he reports its strength[3] and its actual strength will be judged together. Shall we argue that, since he seems to be sincere when he says that his favour is strong, it must be strong, or that, since it does not seem possible that it is

[1] But there is the odd case in which his favour is now weak, but he thinks that it will become stronger.

[2] But this belief needs considerable qualification. Much depends on the way in which he assesses the present strength of his favour. See pp. 218 and 219.

[3] This is not the same thing as his sincerity when he says 'I will do *A*'.

strong, he cannot be sincere? He himself will nearly always be in a position to distinguish between these two possibilities. But we are like astronomers trying to plot the positions of two stars from the gravitational effects of the binary system. If, impressed by the agent's apparent sincerity and influenced by our general knowledge of his preferences, we chose the first alternative, we would be admitting that our first attempt at a list of escape-clauses yielded a generalization which was not analytic, but contingent and false. However, we could still maintain that there was an unspecific analytic statement connecting strength of favour, judged by criteria independent of the sequel, with the sequel: i.e. we could still say that it was analytic that there must be some further factor, as yet undiscovered, which would explain why the action was not performed. But then we should have to try to discover what that factor was. Certainly we would not accept the situation with equanimity.[1] Even more certainly the concept of favour, in its present form, could not survive a general dissociation from action. But from none of this does it follow, nor is it true that, when the agent says 'I will do *A*', he cannot be sincere if his intention is inefficacious.

My use of the concept of favour is, admittedly, a distortion, because favour is not sufficiently basic or generic,[2] and an over-simplification, because the different species in the genus are built up in various complicated ways. But it is sometimes legitimate to ignore the finished surface of our conceptual system, and I hope that, though I have not isolated the basic generic concept that I mean, I have at least indicated it. It ought to be possible to take it as a foundation, and, using such things as intensity, belief and agency, to reconstruct our present system in a way that would show in detail how inductive and imme-

[1] There is an element of rather transparent bluff in the phrase 'unspecific analytic statement'. It suggests that in a particular case, when the agent did not perform the action and none of the escape-clauses applied, we could use the analytic statement as a premiss and argue that his favour could not have been strong. But how could we ever verify that there was no further factor operating? Struck by this doubt, we might reverse the argument and, convinced on other grounds that his favour was strong, infer that there must be some further factor operating, which ought therefore to be added to our list.

Even when such a crisis has not yet occurred, it might occur at any time. Before it occurs, it is possible to treat the list of escape-clauses as if it were complete, and to base the concept of strength of favour on two working criteria which may not continue to be satisfied together. Saying that their joint satisfaction must continue would be like saying that there must be a point at which two lines meet, although it is not geometrically necessary.

[2] What is?

diate knowledge fit together. For instance, deprive an agent of any confidence that he will continue to favour a project, and what is left of his decision to carry it out is only his present favour.

There remains to be considered the other transition where it is difficult to see how inductive and immediate knowledge fit together, the transition from desire through deliberation to selective favour, and, perhaps, decision. Here the question is what is the relationship between the agent's immediate knowledge, based on deliberation, of his own selective favour, and perhaps decision, and the prediction that he would favour that particular project most and perhaps decide on it, which was based on inductive reasoning. Now we have seen what happens when he tries to combine both things. His prediction in some cases alters to a certain extent the nature of what he predicted. But when they are not combined, the fit ought to be exact, provided that the prediction goes through the pattern of his desires, and there is no rationalization. How does this exact fit come about? At first sight it looks as if nothing could be simpler. The agent just does the things that were predicted of him, and, though he is unaware of the prediction, he knows that he is doing them. The prediction was based on knowledge of his desires, and his decision, if there is one, is based on the same desires, of which he will usually be aware. But it is not so easy to see how the decision is based on the desires, or how the same desires can be known in two different ways.

How is the decision based on the desires? What is acting for a reason? It could hardly fail to be true that the agent's reason for his decision or action is a cause in some sense of that versatile word. The difficulty begins when we try to fix the sense in which it is true. On the one hand, Wittgensteinian discussions of this problem fail to do justice to the force of the word 'because' in 'I did it because . . .';[1] and, on the other hand, those who do justice to its force usually understate the differences between reasons for decisions or actions and other kinds of causes. The problem is notoriously difficult, and it is easier to say how it will not be solved than to say how it will be. For instance, the idea that a solution might be extracted from the analogy between intentions and commands cannot be right, if only because issuing and obeying commands are both instances of the very thing that requires to be explained.

I shall approach the problem by first trying to answer the

[1] This is argued by Professor Davidson in his article 'Actions, Reasons and Causes' in the *Journal of Philosophy* 1963.

other question: How can the same desires be known in two different ways? Now the way which is open to agent and spectator alike strongly suggests a dispositional analysis of desire. It is, however, well known that this analysis sometimes errs by over-simplification and omission. It over-simplifies by assimilating ascriptions of reasons for actions to ascriptions of traits of character. What it omits is the way of knowledge that is open only to the agent.

It is not difficult to begin to correct the over-simplification. The first step is to distinguish between a desire to do a particular thing and a general desire. Now a desire to do a particular thing may be blocked either by another such desire, or by an impediment that is not a desire at all. But, even if it is blocked, its existence must make some difference to the world. Perhaps it will show itself at some other point in the person's behaviour. At least, the prevailing desire should be implemented with some reluctance, or the impediment should produce some frustration. Alternatively, or perhaps in addition, we may require that he should have, and show by his behaviour that he has, a general desire that gives some support to the rejected project. If so, it does not matter that often there will be no name of a trait of character associated with the general desire. All that is needed for this version of the dispositional analysis is that the kind of thing that is desired should be identifiable, and that the agent's desire for it should be manifested in his behaviour. If the strength of his desire and its manifestation in his behaviour were below the average, as they might be, the relevant trait of character, even if it had a name, would not be ascribed to him. The dispositional analysis of desire is not so closely tied to the dispositional analysis of traits of character or even of moods.

But, however much we elaborate this version of the dispositional analysis of desire, it does not allow for the way of knowledge that is open only to the agent. In order to correct this omission, we must add the kind of thing that I mentioned in my account of deliberation: we must say that a desire is a tendency not only to behavioural manifestations but also to inward favourable reactions. It is quite absurd to neglect the pervasive influence of desires on thoughts and fantasies.

The concept of a reaction does not cover the whole of this wide field. For a favourable reaction to an idea is not the only kind of inward effect of a desire. Desires do not merely add colour to our thoughts and fantasies: they also exert an influence on their structure, and even on their existence. The

word 'reaction' suggests a definite situation, in which something is presented for assessment. That is why it is appropriate when the agent is deliberating, and in practical matters it applies both to the outward and to the inward effects of desires. But, for that very reason, it fails to cover the whole field of the inward influence of desire.

However, the fact that human beings have inward favourable reactions supplies most of one part of the explanation of the way of knowing desires which is open only to the agent. The other part of the explanation is supplied by a fact which much of this lecture has been devoted to emphasizing, the fact that in most, but not all people these reactions are on the whole, but not always consistent with each other and with behaviour. Of course, a reaction of this kind does not have to occur every time that anyone acts, or even decides to act. Nor, when it does occur does the agent need to remember earlier reactions in the same set, as he would, if he were carefully testing a thing for a dispositional property. These reactions point beyond themselves: that is to say his inductive assumption, that they are on the whole constant, is confirmed.

These things might have been otherwise. Human beings might not have been capable of thought. Or they might not have been susceptible to any kind of pleasure or pain. In either case they would have had no inward reactions of favour. Or, although they had both these capacities, they might have been incapable of action. In that case the concept of favour would have developed differently: it might perhaps have produced the concept of wishing, but not the concept of intending or the other concepts in that family. But, given all three things, there is the possibility that there should be, on the whole, consistency between the inward favourable reactions of one person, consistency in his behaviour, and consistency between his inward favourable reactions and his behaviour. This possibility is realized for most people, and it explains our concept of favour and the way of knowing desires which is open to the agent.

This brief sketch of the way in which the two kinds of knowledge fit together at this point leaves many problems unsolved. For instance, there is the fundamental question how the agent knows what the object of his present favourable reaction is. There is also the question, how he knows that his reaction to a particular project is more favourable than his reaction to its rivals. So far, I have discussed only the possibility that he might be mistaken when he predicts that he will continue to favour

most the project that he favours most at the moment. But can he be mistaken when he says that he favours it most at the moment? Or, to make things even easier for him, can he be mistaken when he says that he favours it to some extent at the moment? Here, at least, he seems to be infallible..

But suppose that he is asked why he favours it, and that he answers that he just does favour it under the description already given, or else gives further descriptions under which he says that he favours it. Can his present reaction to the features picked out by his descriptions be favourable even if none of them are connected with the general pattern of his desires? To put this question in the form which is required by the extended version of the dispositional analysis, can his present reaction to the features that he gives be favourable even if none of them are believed by him to be, or to be connected with a kind of thing desire for which has made, or will make some difference either to his inward life or to his behaviour? If not, could he think that his reaction was favourable when it was not?

These are marginal questions. They suggest things that are on the brink of conceptual impossibility. If we give them a negative answer, our position might be that it is psychologically impossible for him to be mistaken when he reports that his present reaction is favourable, and that, if *per impossible* he were mistaken, we could rely on another psychological impossibility, the impossibility of a totally aberrant reaction, and correct him. Can we credit him with the same infallibility when he says which of two projects he favours most at the moment? If we do, the position will be different. For if *per impossible* he did make a mistake when he said that at the moment he favoured a particular project most, we should not always be able to correct him, since, however well we may know a person, we never possess a system of reliable general statements assigning precise and unvarying relative strengths to his desires.

But is it really psychologically impossible for a person to be mistaken when he says which of two projects he favours most at the moment? Perhaps it is, even when the predicament is very delicately balanced. But this infallibility would be secured only by restricting his report to the present moment. When what he reports is less restricted in time he can certainly make a mistake. Moreover, there is another connected way of assessing present favour which is certainly liable to error, even when it is restricted in time. Let us apply the word 'par' to the degree of favour which is just sufficient to produce the following result:

if he were now given the opportunity to perform the action, provided that there were no impediments except other desires, he would do so. Then if he reports that his present favour is at least par, he may well be mistaken.

It would take too long to complete this sketch of the way in which the two kinds of knowledge fit together at this point, but I would like to defend it against one criticism. The suggestion that desire is partly an inward reaction of favour might be criticized for assimilating a desire, to that extent, to a sensation, and for implying that the agent has to recognize it in the same way that he has to recognize a sensation as one that accompanies a bodily need. But there is no such assimilation. For he does not start from the reaction, and ask himself what its object is: the direction of fit is the opposite one, so that, by the time that he has the reaction, it is already distinguished from others by its object, and his problem is to assess its strength. Of course, this does not dispose of the fundamental question, how he knows that the thing which seems to be the object of his reaction really is its object. My present point is only that the question which he asks himself is not whether it is the object of his reaction, but, rather, how strong his reaction to it is.

Finally, there is the question how the agent's decision is related to his desires. I have implied that this question is likely to get the same answer as the question how his action is related to his desires, and that the answer would provide an analysis of the concept of acting for a reason. In the discussion that follows I shall concentrate on this concept. What I shall say about it can be generalized, without much modification, to cover the concept of deciding for a reason. I have already suggested that the answer could hardly fail to be that the relationship between desire and action is, in some sense, causal. But in what sense? Is some general statement implied by the statement that a person did something because he believed it to be, or to be connected with something that he desired?

It has been pointed out that the statement, that *A* caused *B*, does not imply that *A*, under the description '*A*' was a sufficient condition of *B*, under the description '*B*'; and so does not imply the straightforward generalization that whenever *A* occurs, *B* ensues: it only implies that *A* and *B* fall under some descriptions, perhaps as yet unknown, which would yield a true generalization of that form.[1] Now those who say that a reason for an action is a kind of cause usually mean at least that the agent's

[1] By Professor Davidson, loc. cit.

desires and beliefs about the situation caused his action. It is then natural for them to assume that, if they are going to give the word 'cause' anything like its usual meaning, they must say that the statement that he did something for a particular reason implies a generalization of the following form: Whenever the state of desire and belief, as described in the singular statement, recurs, the action, as described in the singular statement, ensues. But this assumption is said to be mistaken.[1] The singular statement only implies that there is some true generalization satisfied by the particular case. It might even be a neurological generalization.

It is true and important that an ordinary singular causal statement does not imply its own straightforward generalization. If we knew that the only true generalization which was satisfied by the particular case contained terms which were totally unknown to the person who made the singular statement, we would not deny either that his singular statement was true or that it was causal. So, if what he says is that a person did something for a particular reason, and if the case satisfies only one true generalization, and that one is neurological, his singular statement may well be both true and causal. We could not even object that the agent must have made his decision under a rationalizing description. For that would mean that his action had not really issued from the stated desire.[2]

However, it is also important that, when someone wants to know another person's reasons for his action and cannot ask him, he is not necessarily reduced to guessing them, and that the evidence for such ascriptions is not neurological but psychological. How does the spectator use this evidence? There are two distinct steps in his ascription. First, he argues from his knowledge of the agent and his situation to the present state of his desires and beliefs. Then he argues that the present state of his desires and beliefs sufficiently explains the action. He might generalize each of these steps. He would generalize the first step by main-

[1] By Professor Davidson, loc. cit.

[2] But I suppose that we might regard the desire as epiphenomenal. The assessment of epiphenomenalism turns on the answer to a question which will not be discussed: if a desire is always accompanied by a neurological state, how can we determine which of the two is causally efficacious? In any case, there must be some kind of general connexion between the desire and the neurological state. Otherwise, the singular statement would be a pure guess (not because the speaker would be unaware of the general connexion, but because it would not exist).

taining that, whenever this agent is in a similar situation, that state of desires and beliefs will recur in him. He would generalize the second step by maintaining that, whenever the state of desires and beliefs recurs in him, he will perform a similar action unless there is some specifiable factor to explain why he will not.[1] Of course, neither of these two generalizations will be what is usually called a 'psychological law'.

Let us examine the second generalization, and, in order to avoid confusion between it and the first one, let us assume for the moment that the agent himself is giving his reasons for his action after he has performed it. Then it may be objected that he is not committed to the truth of a generalization of the second type, since all that he needs to do is to give the main desire and belief from which his action issued. But such explanations are often very incomplete, and, the nearer they approach to completeness, the greater will be the certainty which anyone, spectator or agent, would have been justified in feeling, if he had predicted the action, inductively or immediately, from the state of desires and beliefs before it was performed. Suppose, for instance, that the main desire was opposed by another desire, and that it prevailed only because it was reinforced by a third desire. In that case anyone who only mentioned the main desire would not have given a complete explanation of the action, and, if he had predicted it before it was performed, he ought to have felt correspondingly uncertain of his prediction. If, on the other hand, the main desire prevailed without reinforcement, the same explanation would have been more complete, and, correspondingly, anyone who predicted it ought to have felt more certain of his prediction.

Most ascriptions of reasons for actions are meant only as incomplete explanations. There is nothing wrong with that. Nevertheless, we often can, and sometimes do make such explanations more complete. Could we make them absolutely complete? Or is the concept of an absolutely complete account of an agent's reasons for his action a Kantian idea of reason?

These questions are unanswerable without a criterion of absolute completeness. A deterministic criterion would be this: an account of an agent's reasons for his action is absolutely complete if and only if anyone who gives it ought to have felt absolutely certain if he had used the same state of desires and beliefs in order to predict the action before it was performed. But this

[1] This generalization need not be interpreted as an analytic statement. See p. 223, footnote 1.

criterion is hopelessly unrealistic. For even the most elaborate and accurate account of the present pattern of the agent's desires and beliefs could not yield the degree of inductive certainty that is often justifiable when the prediction is about something that is not human. Various sources of uncertainty about predictions of human actions have already been mentioned. There might be an external impediment, or a change in the circumstances which produced a change of mind. Suppose, then, that we say that an absolutely complete account of his reasons will be one based on a state of desires and beliefs which would have given the highest degree of inductive certainty to the prediction that, if there were no such intervention from outside, he would perform the action. But even this is unrealistic. For, quite apart from intervention from outside, a person will not necessarily carry out the project that he favours most at the moment, even if he can do so, and knows that he can. There are also psychological impediments which are not opposing desires. So an absolutely complete account of an agent's reasons for an action would give the highest inductive certainty only to the prediction that he would perform the action, if there were no intervention from outside, and no psychological impediments of that kind. In short, this concept of absolute completeness only covers the system of the present desires and beliefs of the agent, and is not affected by adverse factors outside that system.[1]

Is anyone even then ever in a position to apply this concept to a particular account of an agent's reasons for his action? First, let us consider the spectator. If he cannot question the agent about the state of his desires and beliefs before the action, he will have to use a generalization of the first of the two types that I distinguished just now. But, even if we waive any difficulties about beliefs, generalizations of that type do not yield absolutely certain inferences of other people's desires. For, however well we may know a person, it does not seem possible to find reliable generalizations assigning precise and unvarying relative strengths to his desires. If this is impossible, there are many explanations of the impossibility. Even if people did not change, there is a limit to dependable discrimination in the vast field of objects of desire and aversion. But they do change, not only durably,

[1] Of course we might use a non-deterministic criterion of 'absolute completeness'. We might simply say that an account of an agent's reasons for his action is absolutely complete if it covers all the desires that were at work, whether or not it yields a prediction with the highest inductive certainty. See p. 225.

but also momentarily and capriciously. Moreover, a person is aware of the pattern of his desires, and this too produces unpredictability in various complicated ways. Consequently, the spectator is often unable to collect the data from which an absolutely complete account of the agent's reasons would have to be extracted.

But perhaps the agent is in a better position to apply the concept of absolute completeness to his own account of his reasons for his action. Certainly at the moment of action he does not use the spectator's first generalization, since he knows immediately the state of desires and beliefs from which his action issues: nor is he committed to the truth of that generalization, since, given that favour can change from time to time, its falsehood is compatible with the truth of his account of his reasons. This removes one doubt that is often expressed about the thesis that I am advocating. It is supposed that, according to this thesis, the agent is at least committed to a generalization of the first type. But there is no such suggestion. It is only maintained that he shares the spectator's commitment to the second generalization, or rather to the following more fully developed form of it. Whenever the same state of desires and beliefs recurs in him, he will perform a similar action, unless there is some specifiable intervention from outside, or some psychological impediment that is not a desire.[1] This is a very different commitment. It does not even imply that he really wanted to perform the action. For in the crucial period immediately before the action his reaction[2]

[1] This generalization need not be interpreted as an analytic statement. It would be analytic if, whenever a similar state of belief recurred in the agent, without any intervention from outside or any specifiable psychological impediment that was not a desire, the performance of the action were made a necessary condition of the recurrence of an equally strong desire for the same object. But, though object and strength of desire must sometimes be judged in this way, they need not always be. The agent himself is in a position to report them immediately without waiting until the moment for action arrives: and others can judge them from their general knowledge of his preferences, reinforced by his own report, which is unlikely to be mistaken, and which may be judged sincere by things that are independent of the sequel. Then the generalization will be a contingent statement.

The concept of strength of desire or favour is based on the high frequency with which such contingent general statements are satisfied by particular cases. But the concept conceals its own foundations by appropriating as working criteria the two things which are only contingently corrected. See p. 214.

[2] For the sake of simplicity, I sometimes speak as if he would always have a reaction at that time. But since desire is dispositional, it may only be true that he would have had one if he had paused to reflect.

may not agree with his true and settled preference. If, by using his advantageous position he can surprise a well-informed spectator, that may well be because his reaction is rather aberrant, so that his account of the desires that he felt in that particular situation will be less rich in implications about the rest of his life, and ought not to be used as evidence for a generalization of the first type.

Is his account of his reasons ever absolutely complete? Those who think that it can be must believe that he can sometimes survey the whole field of his desires and accurately assess their contributions at least at that moment. For, whenever he fails to do this, the generalization of the second type to which he will be committed will almost certainly be false.[1] Now there certainly are occasions when his survey of his desires or his assessment of their contribution will be deficient. For instance, he could hardly allow for the existence of unconscious desires. The only way to circumvent this difficulty would seem to be to deny that influences and impediments of that kind are properly called 'desires'. But this denial only reduces the importance of an absolutely complete account of reasons for an action by reducing the scope of the system of the agent's desires: so-called 'unconscious desires' would then have to be treated as enormously important psychological influences and impediments of the other kind. Moreover, even if no desires are omitted from his survey, his assessment of their contribution may well be mistaken. Everyone is familiar with the unnerving effect of this kind of mistake: one's assessment of the contribution of different desires to one's action in one situation does not square with one's actual desires in another slightly different situation; for, according to the original assessment, the slight difference in the situation ought not to have produced such a big difference in one's desires.

However, he may sometimes be able to produce an absolutely complete account of his reasons for an action. And, even if he never could, this concept of absolute completeness would still function as an idea of reason. It would be a point which made the structure of our conceptual system intelligible, even though it lay outside it. For our use of incomplete explanations based on states of desire and belief depends on the possibility of making them at least more complete. Perhaps we shall never discover an absolutely true generalization satisfied by a particular

[1] But there is the odd case in which he overlooks two desires which exactly cancel one another out.

instance of desire and belief and action[1] unless we abandon psychology and adopt some entirely different system of descriptions. But it would not follow that an ascription of a reason for an action implies only that this instance satisfies some true generalization which may even be neurological. For the implied generalization can still be a psychological generalization about the agent's desires and beliefs, provided that we do not insist that it should be true without exception. Perhaps the exceptions will only be the result of our ignorance. For the pattern of a person's desires is very complex and difficult to make out, even if the person is oneself. Alternatively, it may be that the exceptions will be due to the nature of things. For even an account which covered all the desires that contributed to a particular action might still not be complete by the deterministic criterion. Whatever the explanation, it seems likely that this part of psychology will yield only approximate results.

If, deterred by the approximate character of these results, we maintain that the implied generalizations cannot be psychological, or at least cannot belong to this part of psychology, we shall be opening up a gap between the implications of these singular causal statements and the evidence on which they are based. But is such a gap objectionable? Ordinary singular causal statements are often based on very rough evidence, and the person who makes one often has no inkling of the precise scientific generalization which could sometimes be included in the disjunction of generalizations which might be said to be implied by his statement. Moreover, psychological generalizations of the kind that are being discussed may not be necessarily approximate, and so they too may be included in the disjunctive implicate.

However, there are objections to this view. The disjunctive implicate is not fully specified, and therefore not falsifiable. Perhaps, if we did not insist that the generalizations should be true without exception, we might find a type of approximate generalization which could plausibly be said to be implied by singular ascriptions of reasons for actions. Now psychological generalizations of the kind that are being discussed may be necessarily approximate, and therefore not subject to knock-out falsification. But it is possible to collect overwhelming evidence against them. When this happens, it does sometimes lead to the retractation of

[1] It is important to remember that the generalization would include the proviso that there must be no interventions from outside and no psychological impediments that are not desires. See p. 223.

a singular ascription of a reason.[1] Therefore, there is a strong case for saying that such psychological generalizations are implied by singular ascriptions of reasons, and are not merely included in a disjunctive implicate.

There is also another consideration which suggests that, whatever else a singular ascription of a reason implies, it must at least imply a psychological generalization of this kind. If the agent's state falls under a scientific description which yields a generalization which is true without exception, then there must be a general connexion between the scientific description of his state and the ordinary description of it.[2] Now it is hard to see how this general connexion could be established unless the ordinary description of his state could be elaborated until it yielded an explanation of his action which was at least more complete.[3] For, if he only gives the main desire which produced his action, it would be extravagant to hope for a straightforward general connexion between his description of his state and some scientific description of it. That would be like hoping for a straightforward general connexion between the concepts of mechanics and the description of a single causal factor selected from the total cause of a particular physical event.

If we look back at the transition from desire through deliberation to selective favour, and perhaps to decision and action, it will be evident that the requirement that, when there are two ways of knowing that a statement is true, both built into its meaning, we must have some guarantee that they fit together, is met at every point. The spectator uses a generalization of the first type to infer the agent's desires and perhaps his decision in a particular situation, and the agent knows them immediately. The spectator uses a generalization of the second type to infer that the agent will perform the action, and the agent knows immediately that he will. At both these points we can rely on the fit between inductive and immediate knowledge. If, as I have been arguing, all the points in the transition are connected causally, then, if there were two ways of establishing the causal connexions in particular cases, we would have to have some guarantee that they fit together.

This lecture has been devoted to describing and explaining the way in which inductive and immediate knowledge fit together at

[1] See p. 224. The retractation mentioned there is retractation by the agent because he was mistaken. There is also retractation by the agent because he was insincere, and retractation by the spectator because he was mistaken.

[2] See p. 220, footnote 2.

[3] See p. 221.

various points in the transition to action. The fit is very close and complex. Even when the agent takes the way of knowledge that is open to himself alone, he is depending on inductive psychological assumptions which figure as explicit premisses in the spectator's account. So, to revert to my opening topic, when the agent actually uses them as premisses, and predicts his own decision or action inductively, what is happening is that the underlying structure of the conceptual system of self-knowledge is showing through its finished surface.

'Will, therefore, is the last appetite in deliberating.'[1]

[1] Hobbes, *Leviathan*, pt. i, c. 6.