THE VARIETY OF RATIONALITY

Adam Morton and David Holdcroft

I—Adam Morton

My aim is to argue for a claim about the way in which two families of concepts, one clustered around that of intentional action and the other clustered around that of rationality, are related. *That* they are related cannot be a very startling claim; one can easily come up with a number of sketchy formulas to relate them (for example 'only intentional actions can be rational'). The claim I shall eventually formulate entails first that most such easy connections are false or radically ambiguous, and second that the only way in which we can understand the actual conceptual connections here is to take account of the fact that we are dealing with families of concepts rather than single fixed ideas. This second claim, which may well remind you of Austin's dismemberment of the concept of intentional action, is a special case of something which I think characteristic of commonsense psychological terms, that they come in categories, and the conceptual connections are between the categories rather than between the particular members of them that fairly accidental features of our culture present us with. I find it hard to express this thought in terms that are both general and clear, and so one of the secondary aims of the paper is to make sense of it in this special case.

I

Intentions Begin with intentional action. Austin's attack on the idea that we have a single category of things which people do, quite naturally described as 'intentional', which is at the same time the central example of actions in general and a prerequisite for the application of various important predicates to actions, seems now to have pretty well disappeared from influence. And there is a simple reason for this, Donald Davidson. Davidson's way of talking about actions, making their intentionality relative to a specification of an intention, handles some of the intuitions which Austin was exploiting, and does so more smoothly and flexibly than Austin could.[1] Consider an example, Austin's own:

> Suppose I tie a string across a stairhead. A fragile relative . . . trips over it, falls, and perishes. Should we ask whether I tied the string there intentionally? Well, but it's hard to see how I could have done such a thing unintentionally. You don't do that sort of thing—by accident? By mistake? Inadvertently? On the other hand, would I be bound to admit I did it 'on purpose' or 'purposely'? That has an ugly sound. . . . Maybe I had better claim I was simply passing the time, playing cat's cradle, practising tying knots.

I take Austin's point to be that a claim that tying the string was unintentional will not do as an excuse because it is so clearly false. So if I want to evade responsibility for killing my relative I had better use some other style of excuse. But of course the natural thing to say is that one *had* tied the string intentionally, perhaps in order to play cat's cradle, but that one had not intentionally killed the relative. That was done by mistake. The two acts can even be identical, as long as what is done is intentional as described one way and unintentional as described another. This is the way Davidson would have us describe the case, and what could be simpler?

It *is* enlightening to talk of actions as intentional under or relative to descriptions. Does that show, though, that there is just one important division of actions, into intentional and non-intentional? One price we pay for doing things Davidson's way is the deviant causal chain problem. The connection is as follows: An action is intentional if and only if it is caused by the intention under which it is intentional. As many familiar examples show us, not all causal connections between intentions and actions satisfying them qualify those actions as intentional. There are deviant causal chains. So there must be a right kind of causal connection, and since in all cases the relevant factors are the same—intention, causal link, and resulting event—the analysis of 'right kind' must be the same in all cases. So we must try to find a single set of conditions which will in all cases dis-

[1] J. L. Austin 'A Plea for Excuses', in *Philosophical Papers* Oxford: Oxford University Press, 1970. Donald Davidson 'Agency', in *Agent, Action, and Reason*, R. Binkley *et al* eds., Toronto: University of Toronto Press, 1971.

criminate those causal chains which do from those which do not generate intentional actions.

We inherit this project if we see things Davidson's way. Notice how different the possibilities seem if we stick with Austin. There are then a number of different subcategories of action, and what we are to determine of an action is not whether it is intentional but to which of these subcategories it belongs. The criteria of membership may vary from category to category in many ways, including that of the appropriate causal connection. A causal connection that could establish an action as, say, done on purpose, might not show that it was done deliberately.

The question then is: is the deviant causal chain problem—or, rather, the examples which seem to suggest such a problem—best handled by treating intentional actions as being all of one kind, or by dividing them into sub-cases? (And if the latter, do the sub-cases correspond either to Austin's intuitive labels or to anything systematic in the ways we describe and explain action?) It is the sort of question that is naturally attacked by making plausible analyses, testing them against examples, revising them, and trying to see the general resulting direction. But by the nature of the question—since it is an inquiry into the solidity of the concept of intentional action and the roots of intuitions involving it—we cannot use as raw material just our reactions to putative cases of intentional action, our dispositions to apply or withhold the label 'intentional'. We need to frame examples in other terms, which carry less danger of begging the question. Austin's discussion of examples of action is nearly always linked to blame or moral responsibility in some way. I think that this is not in fact a promising way to begin, for the link between any of these categories of action and any kind of moral evaluation is just too loose. Too many other factors can get in the way, so that there are for example no end of examples of intentional actions with foreseen bad results for which the agent bears no blame, and pretty clear examples of unintentional actions, accidents in fact, for which one is responsible.

We must start with a rough idea of what the function of classifying events into intentional and non-intentional, or in similar more complicated ways, is. One function, at any rate, seems pretty clear. We explain actions in terms of states and qualities we attribute to people, and we make these attributions largely as inferences to explanations of their actions. The reasoning from act to state or vice versa—the style of explanation linking character, intention, skill, belief, desire, and the rest, to action—can be very complicated. But some crude categorizing principles provide short cuts and warn of dead ends; if you did it by mistake then you could not have done it out of cruelty; if you did it deliberately then there was skill involved. These principles make membership in a subcategory of action a necessary (or, occasionally, a sufficient) condition for an action to be explainable in a particular way.

What we need to do, then, is to look at the ways in which actions can be classified into kinds in terms of the ways they can be explained. We need to use crude intuitions of the similarity of explanations, weigh them against analyses of psychological explanation, and see if the result sheds any light either on the concept of intentional action or on the variety of causal link which can connect intentions and actions.

II

Accomplishments and Successes Here are some typical examples of members of two broad classes of action, which contrast in various ways with traditional standard cases of intentional action. In describing them I try to describe the causal connection between the action and the motives and traits of the agent, because that will eventually connect back with questions of the normality of causal chains and connect forward with questions about rationality. In what follows I shall use 'action' in what may be an unnaturally wide sense, covering anything that agents do, reserving 'act' for a narrower class more like that sometimes covered by 'action'.

A: Accomplishments[2]

(*i*) James Joyce remarked, some time in the 1920's, that T. S. Eliot had made English poetry an occupation for grown-up men, that he had rescued it from being an effeminate quasi-infantile pursuit. Let us suppose that Joyce was right. He was describing an accomplishment of Eliot's over a period of some ten years,

[2] In an earlier draft I called Accomplishments 'Harrisonian actions' because of their connections with the themes of Andrew Harrison's *Making and Thinking* (Harvester Press 1979.)

describing it in terms that might not have, perhaps could not have, occured to Eliot. But the description is not just a description of an incidental effect of Eliot's actions. It captures an essential feature of what Eliot was doing. For during this time Eliot's poetic intentions evolved in a certain direction, with a certain rationale. His poetry was the result of this development of his intentions; and Joyce's remark captures this fact. Joyce's description thus says what Eliot accomplished, what we must credit to him, although it does not correspond to anything Eliot set out to do.

(*ii*) I set out to make a student feel small. I begin by berating him about his terrible essay. As I develop my invective I become aware of my victim's fears of his family's reaction to his academic incompetence. On impulse, I allude to a nonexistent acquaintance with his father's books. The trick works, and soon, by reacting skillfully and luckily to his reactions I produce complete humiliation and misery. I'm a swine; a great range of attributes apply, and can be used to help explain my action. They apply in virtue of my exploitation of hunches and chance events, in pursuit of an aim that only became definite as I achieved it.

B: Successes

(*i*) I am locked in a room from which I can escape only by singing an accurate E-flat. (That will set off a mechanism keyed to release me when the church bell strikes, say, when it will be too late to prevent my loved one's abduction.) I cannot sing in tune, let alone produce named notes. Too many incorrect tones will jam the mechanism, making it refuse to release me. Frantically, I search my memory for the sound of the church bell, think of a certain black key on the piano, touch wood, and sing. It works. I escape (and rescue my beloved). Later, on being congratulated, I protest, truthfully, that it was mostly luck. But, still, I did do what I meant to, in the way I meant to: I tried something and it worked.

(*ii*) Again I am locked in a room and can only get out by sounding E-flat. There is a piano in the room, but I cannot play the piano, cannot even name the keys. The locking mechanism —this time—will not jam on receiving the wrong tone. So I strike all the keys on the piano. Several of them are E-flat, the door unlocks, and I escape. Again I did what I meant to, what I tried worked. No luck is involved this time. But, still, there is a certain lack of specificity in my production of that E-flat: I didn't mean *that* key to produce that note.

None of these are straightfoward cases of intentional action. They fall short of what might be required in either of two ways. One—seen in the Accomplishments—is a lack of specific intention. Eliot determinedly, brilliantly, admirably set English poetry on a new course. But he may never have said 'I'm out to set English poetry on a new course'. For all that, what he did was not done out of ignorance or by mistake; there was a definite guiding intelligence at work, which need not have summed itself up in an intention. And in the humiliation case [A*ii*], I do not set out with the specific intention of producing the specific humiliation that results, but the specific humiliation, down to quite fine details, is attributable to me, because I guided the situation towards it, with a developing intention that became more specific as the episode took shape.

With Accomplishments, then, the presence of a guiding, evolving control of some aspect of things allows both the presence of accidents and the lack of specificity of intention. One stringency compensates for the lack of another. With Successes, on the other hand, the stringency that is satisfied is the fitting of intention and result. This compensates for a lack of causal determination between intention and action; a quite different result could easily have been produced.

III

Intentions and Descriptions In both kinds of case it seems intuitively clear that something about what is done is intentional and something is not. We might expect to be able to express this in Davidson's way, as a matter of the descriptions under which the acts were or were not intentional. Thus in B*ii* [Random Piano Keys] the act clearly is intentional under the descriptions

—unlocking the door
—pushing that key (which happens to be E-flat)

and is not clearly non-intentional under the description

—producing E-flat

since that is what I had set out to do. The only description under which it is *clearly* not intentional is

—producing E-flat by pushing that key.

Case B*i* [Desperate singing] is similar. While the action is clearly intentional under the descriptions

—unlocking the door
—doing *that* with the vocal apparatus
—producing E-flat by doing *that* with the vocal apparatus

and not clearly non-intentional under the descriptions

—producing E-flat
—producing E-flat by singing

the only description under which it is pretty clearly not intentional is

—producing E-flat by aiming at *that* sound

(where the 'that' can only be filled out by producing an E-flat).

The point is the difficulty of finding descriptions under which these actions are clearly not intentional, although there is clearly something non-intentional about them. The only descriptions under which they are uncontroversially non-intentional turn on *de re* references to aspects of the production of the resulting states of affairs ('by pushing that key', 'by aiming at that sound'). That is already an important difference from the Davidsonized Austin case (intentional under the description 'tying a knot', non-intentional under the description 'killing the relative') for it raises a doubt whether the relativity to a description is really just a relativity to an intention (as Davidson intended) or whether it can also indicate some variability in the specification of the connection between intention and result.

For what is it to intend to produce *that* note, if it does not mean—as it does not here—just intending to produce a note satisfying the right description ('E-flat')? It must be either to have an impression of that note or a way of identifying it (for example a disposition to recognize it on hearing) as a guide for the production of it. Similarly, if one intends to produce the note by pushing that key, and the intention is *de re* of the key—one intends of the key that by pushing it one will produce the note—then one can guide one's production of the note by use of one's capacity to refer to it. If the key is referred to via a visual link, one's pressing of it is anticipated as a physical action in the direction-as-seen; if it is referred to more indirectly, e.g. as the key you get to by putting your finger on the dirty white one and extending your middle finger forward and to the right to the nearest black key, then the projected mode of production is different, as indicated. My suggestion is *not* that whenever one intends to do something, and while intending thinks of an aspect of the act or its environment in a *de re* way, one thereby intends to do it in a way that exploits the referential link. It is rather that when we *say* of someone that they intend to do something, describing the intended act in terms of a referential link between the actor and some object or aspect relevant to the act, then we are usually ascribing to the actor a plan of acting with reference to those objects of reference, and by use of the referential tie. And in describing an act as intentional with respect to an intention, we usually mean that it was carried out by use of the connections involved in the referential tie.

It is this, I suggest, that marks off intentions from desires and expectations: we single out some of a person's states as intentions, not because they are different in content or psychological role from others, but because in so labeling them we are signaling a claim that their content specifies the information (better, information link[3]) with reference to which the agent will act. And—tied to this—in saying that an act is intentional, in accord with an intention, we mean that it is guided by the information (link) alluded to in the relevant ascription of intention. So: in saying that an act is intentional according to some stated intention, we are *not* just describing the initiating intention and postulating a somehow standard (normal, nondeviant) causal connection: we are actually specifying in part how that connection proceeds.

Two consequences. First, sometimes the use of a description of an intention to pick out an intended causal connection will clash with its use to specify the content of the agent's desire. This

[3] See Gareth Evans *The Varieties of Reference*, Oxford: Oxford University Press, 1982, Chs. 5 and 6, for more about information links.

happens with the actions I have called accomplishments. They involve the careful use of information about various aspects of the environment, which we can package together with a description of the result and of the manner in which the agent steered things towards it, by saying what the essential accomplishment was. But then there is no easy connection with definite pre-existing (or sometimes even post-existing) desires.

Second, sometimes the description of the ways in which the action was and was not controlled cannot be incorporated into a statement of the agent's desires-as-intentions, although there is a perfectly definite guiding desire. So we state the desire and say that that was what the agent was up to, and use more subtle means to describe the mode of control. This happens with the actions I have called Successes. There the intention is clear, and it fits the action, but relevant facts about the manner and extent of the agent's control cannot be worked into a specification of the intention. You have to state them separately.

IV

Modes of Control. I have so far made one claim and one hint. I claimed that we describe actions not just in terms of intention and result, but also in terms of the kind of connection which holds between them. Often we manage to describe both at once, and the main reason why we can is that many ends are achieved in standard ways. So when we characterize an action as, e.g. shooting a pistol, we implicitly suggest that something like the usual means by which people manage to shoot pistols has been used—contraction of muscles controlling a finger by the finely-tuned use of the efferent nerves of the central nervous system. (It may not be a finger, but it is expected to be something similar enough to show the same general responsiveness.) Another kind of description, e.g. 'setting English poetry on a new course', can conjure up a quite different set of expectations about the expected mode of control.

The hint was that these ideas about actions and intentions provide an attitude to the deviant causal chain problem. Now there is a simple way in which if what I am saying is right there is no deviant causal chain problem. For if one can specify intention and type of causal connection independently, then one can define any subcategory of action one has a mind to, e.g. 'act of conscious volition linked by infallible causal mechanism to effect exactly satisfying it', or, 'vague ambition somehow forming part of the grounds for a situation making the ambition roughly content'. And any of these are perfectly possible kinds of action.

They are not all perfectly actual concepts of action, though. We certainly do pick on some particular combinations of a type of intention and a type of causal connection, to constitute the sort of deed we are prepared to explain in terms of motives and traits of character. Now to some extent our collection of subcategories of action may be an accident, determined by accidents of our selection of trait and motive vocabulary, so that all that can be done is to list and classify the local varieties. But, still, it is hard to believe that our collection is *purely* accidental, that there is no rationale to it, even if a fairly local rationale.

My attempts in this section to work out some pattern to the various 'natural' combinations of intention and cause form a digression before the second part of the paper (V–VII), which is largely independent of them. I do not doubt that in all familiar kinds of action there is a causal connection between on the one hand a psychological process, of the management of information and the evolution of desire, and on the other hand a result in the world. And I assume that this connection takes the form of what Peacocke (improving an idea of mine[4]) calls differential explanation: roughly, that had the process gone differently in some specific respect then the result would have been different in some corresponding respect. But what the relevant respects are, and what the correspondence between them is, no doubt varies from one category of action to another. (And the *range* of such correlations varies from one style of vernacular psychology, one local conception of mind, to another.) And so this assumption just reformulates the question (and does leave us with a sort of a deviant causal chain problem): what differentially correlates with what?

Here, briefly, are three more categories of action, followed by an attempt at a generalization from my five cases.

[4] Christopher Peacocke *Holistic Explanation*, Oxford: Oxford University Press, 1979; Adam Morton 'Because He Thought He Had Insulted Him' *Journal of Philosophy* 72, 1975, pp. 5-15.

C: Sub-acts.

Consider spontaneous speech. One knows generally what one wants to say, usually, but rarely knows very specifically the content of the particular sentences one produces, let alone the particular words, until they emerge from one's mouth. Let me assume that on occasion the choice of a particular word—the choice between rough synonyms, say—is quite undetermined by the guiding communicative intention. One produces a word, 'pig', say—instead of 'swine' or 'hog' or 'sow'—and thereby performs an action satisfying the description 'saying "pig" '. Now, contra Davidson, it is perfectly natural to call this an intentional action, or at any rate a non-accidental, non-unintentional one. It is not just the act of asserting the whole proposition that is intentional. The choice of that particular word is also something like intentional. For: it was done as part of a deliberate course of action; it was something whose performance one could not just shrug off as an unforeseen result of one's action; if the choice was brilliant or disastrous it would not be out of the question to attribute brilliance or disastrousness to one. And the reason for all these things is pretty clear: the process which produces it, although not determining it, is in control of it in a particular way, such that whatever act is produced, from the foreseen range, the process will take account of it and produce a suitable continuation, and the process has set up the gap into which the act fell, in such a way that such an improvised continuation would be possible. Such actions, whose manner of intentionality is to fit smoothly into the achievement of larger intentions, though not pre-intended themselves, are in a way like miniature Accomplishments. I shall call them sub-acts.[5]

(Two observations: The continuation, or at any rate the potential continuation, of the larger course of action beyond the sub-act is essential. They are Sartrian in this respect: you have to make them have been intentional. And they are not naturally described as done deliberately or on purpose. To say that one had said a particular word deliberately would be to suggest that there was a controlling intention towards that very word.)

[5] My category of sub-acts overlaps with Brian O'Shaughnessy's class of sub-intentional acts, see *The Will* Cambridge: Cambridge University Press, 1980, Vol 2, Ch. 10.

D: Unaccidents.

A student misses my tutorial. His excuse is a flat tyre on his bike. A reasonable excuse; but next week he misses it again, and this time he has an emergency dentist's appointment. Then there is urgent CND business, then a catastrophic flu. All good excuses, but there seems to be a pattern, and when I discover that he never misses a lecture and is known for punctuality I begin to suspect that he is bored or intimidated or upset by my tutorials. It may be all a series of coincidences. But supposing that it is not, *and* that all the excuses are genuine, then the actions of missing my tutorials are examples of what I shall call unaccidents. They are not complete accidents because they can indicate something about the agent's desires and attitudes, and they are not standard actions because they have immediate causes which have nothing to do with their descriptions as unaccidents. The discovery that there is such a class of actions is recent. No doubt the influence of psychoanalysis is in part responsible for our admitting them as explainable in an evolved common sense, and no doubt the thinness of the line between useful commonsensical conjecture and psychobabble should make anyone a little hesitant about this development. But few would doubt that such explanations are sometimes accurate, and that the culture as a whole is muddling its way towards some assimilation of them to its more standard patterns. One easy and not very controversial assimilation is the special case in which the unproblematic causes of the actions could have been overcome or avoided: my student could have got a lift when his tyre was flat, gone to the dentist an hour later, and so on. Then the pattern of action is a pattern of failures to overcome or avoid these individual causes. We can then quite easily see the act as governed by an intention, of which the agent was not aware, which works by exercising a selective influence on the force of various motives and the attention paid to various facts.

E: Acts.

Accomplishments, Successes, Sub-acts, and Unaccidents are all rather different from intentional action as classically conceived. (Where the classics are Anscombe, Chisholm, Danto, and Davidson, summed up in Hornsby,[6] say.) I see this classical

[6] For a bibliography of this tradition see Jennifer Hornsby *Actions*, London: Routledge & Kegan Paul, 1980.

conception as an attempt to bring into one class the whole variety of actions, and I am thus suspicious of it. But there certainly are no end of actions that fit the classical picture. Let me call them Acts. They consist of an event in the physical world occurring in a shortish space of time (less than an hour, say) linked to physical motions of the agent's body by causal connections which the agent intended, and satisfying a desire of the agent which causally initiates the bodily motions. These conditions are not sufficient, but they are true of the usual arm-waving, well-poisoning, girl-kissing examples. They are stringent conditions, not met by acts in my subcategories above. It is not clear what the rewards for meeting the conditions are: actions from the other subcategories can often be plausibly described as intentional or done on purpose, and attributions to the agent's moral character and the like can be made on the basis of them. Most importantly, Acts do not seem to have any distinguished place in the explanation of action: most of the actions we explain differ in one way or another from Acts.

What do Accomplishments, Successes, Sub-acts, Unaccidents, and Acts have in common? This much, I think: in all cases there is something that may be called an overall *intention*, though it may not be known to the agent (Accomplishments, Unaccidents), and may not have as its content the description one would naturally apply to the action (Sub-acts). In all cases this intention is related to a process of acquiring and evolving beliefs, desires, and plans of action, which results in a change in the world, though this process may result in or be summed up in the intention (Accomplishments) as well as being initiated by it (Subacts, Acts, Unaccidents). And in all cases there is some sort of a differential explanation of the change in the world by elements of the process. Different categories of action focus on different kinds of process, different ways of making up your mind and then putting it into effect, and thus require that the resulting change in the world be responsive to these aspects of thought and performance. (Acts correspond to the case in which the agent's main desires do not change in the course of action, are conscious, and result in deliberation about the means to fixed ends.)

Even this sketchy generalization allows us a slightly different view of the usual examples of deviant causal chains. If it is right, then by changing details of an agent's larger intentions and style of agency we should be able to shift the focus from one class of actions to another, and thus transform an example of an action which is non-intentional through deviance of the causal connection into one in which that same connection is sufficient to establish it as belonging to a suitable different category of action. Consider one of the best examples, Davidson's: the agent is holding a rope and someone is hanging on the end. He means to let go and drop the other, but the enormity of the situation makes him clench his hands in a cramp of conscience. Then nervousness and frustration produce excitement, and a rush of adrenalin makes his hands unlock, so the other falls. (It can all happen very quickly, as quickly as normal control of the hands.)

The action in the example is meant to be clearly non-intentional. But consider two variations on it. First suppose that the agent is often afflicted with muscle spasms causing his hands to cramp. Knowing this, as he tries to release his grip on the rope he searches around for means, thinking relaxing thoughts, visualising his hands open, tensing and then not-tensing them. Finally, the last of these succeeds in loosening some of the fingers of one hand. He releases the rest of that hand with his nose and then uses it to tear the other hand off the rope. The rope is released. If the story is told in this way then what the agent has done is a pretty unpuzzling Accomplishment (like the Eliot or the humiliation examples). The agent managed to achieve something fairly complex (under the circumstances) by exploitation of a series of events over a period of time. We can certainly say that the act was done on purpose or deliberately, and with the story as background we can say that it is intentional.

Next suppose that the agent is a phenomenological philosopher, interested in the qualitative feel of acts of will. In fact he has set up the situation just in order to know what it feels like to let go both of a rope and of a human life. He has practiced letting go of ropes with various weights attached, and imagined various people as the weights. Now here he is with a real particular person at the end of the rope, and he does what he had planned, summons *that* qualitative act of will; but it doesn't come, half comes, and the hands stay closed. He curses, forgets his epoche, annoyance and excitement increase, and the hands open. Again we know what to say of the case. It is a Success like the first E-flat case. Or, rather, like that case would be if we had an agent with perfect pitch who loses it in the strain of the moment and has to grope and pray for the note like the rest of us. The agent plans to do *that*, accomplishes something that satisfies a verbal transcription of that, but does not employ the information-link specifically intended. It is clear what is intended, successful, not too misleadingly called intentional, and also what is lucky (not for the person on the other end), and what is unsuccessful.

These two variations on the original man-on-a-rope example are meant to be quite similar to it. They differ in that they are not puzzling: the usual vocabulary of action applies naturally to them. This tells us what it is that is really problematic about the original example, I think. It doesn't have a subcategory to belong to: we cannot easily see how to place it among other actions in terms of the patterns of psychological explanation it calls up. It could happen that some such examples became important to us, that we expanded our psychological vernacular to make room for them, and then used them as bases for unproblematic attributions to their agents. Then they would no longer be deviantly in between the subcategories, but normal examples of their own kinds of action.

V

Kinds of action, kinds of explanation If there are different subcategories of action it is pretty likely that there are different 'styles' of psychological explanation, each focussing primarily on actions of a particular subcategory. If the bulk of the commonsense principles available are of the form 'under conditions C people are likely to do A', where A is an action of some particular type, then that type will be a mark of that style of commonsense. Not a very profound mark, perhaps: it could be due more to accidents of vocabulary than to anything systematic about the patterns of explanation being used. In the remainder of this paper I shall argue that it can be a more profound mark, that we can imagine a family of styles of vernacular psychological explanation, helpfully classified in terms of the kinds of action they take to be the primary objects of explanation.

I shall concentrate on explanatory principles that rationalize the actions they explain, in a way made generally familiar by a number of writers.[7] I shall exploit the freedom obviously allowed within the general pattern '*e* was wanted and *a* was thought to be a means to *e* so *a* was performed'. And I shall try to make use of the main point of the earlier sections of this paper, that we understand action not just in terms of what happened and what was wanted, but also in terms of how the agent brought it about.

Consider two rather different examples. First: an agent wants a Mars bar and believes that the only way to get one is to steal the bar that her husband has packed with their child's lunch. So she does this. To *explain* her action, though, in a satisfactory way, we must answer some more questions. Did she not mind depriving her child of his chocolate? How did she get around her inhibition on stealing from her child? Suitable answers to these questions will tell us why she did what she did, and may also allow us to describe her action as a reasonable thing for her to have done. (Perhaps her need for the Mars bar was desperate, and her child hates them. Perhaps putting one in his lunch was a cruel joke by his father.) But we need the answers before we can either see the action as reasonable or think we have explained it.

The conclusion I would draw from the example is that both explanation-by-motive and attributions of rationality require a description of the *dynamics* of the agent's desires. We need to know why getting a Mars bar continued to be one of the agent's priorities even after she realized that the only way to satisfy it was to rob her child. The dynamics of desire are even more obviously relevant to the explanation of Accomplishments, as in my second example, which is just example A*ii* above, in which I humiliate a student by suitable emphasis on his academic weakness. In explaining what I did here, it is necessary not just to provide an account of why I set about assaulting the student's self-respect—that much is like the example of the paragraph above—but also to explain why I followed the particular form

[7] See Lennard Nordenfelt *Explanation of Human Actions*, Uppsala: University of Uppsala Philosophical Studies, 1974, and Colin McGinn 'Action and its explanation', in *Philosophical Problems in Psychology*, Neil Bolton, ed., London: Methuen, 1979.

and strategy of assault that I did. This is not something that resulted from a particular piece of practical reasoning, but is something that developed over a period of time (half an hour, say) as I became aware of the possibilities. My aims developed: I acquired some as I came to see that they could be realized, others I acquired as my grasp of developments allowed me to formulate them, still others as means to original or newly acquired ends occured to me. To say this is not to give more than the very beginning of an explanation of the evolution of the aim. (A large part of that explanation would presumably appeal not to antecedent desires but to the rotten details of 'my' character.) But it gives part of what is required, by beginning an explanation of the manner in which I pulled the accomplishment together. And it does this by indicating the dynamics, the patterns of changes, of my desires.

Accomplishments differ from 'smaller' actions in that it is obvious that in explaining them, and in evaluating their rationality, we have to consider changes in agents' systems of desires, and not just their content at any given time. This is a difference only of what is obvious, though, and not of what is true. For *any* kind of explanation-by-motive will require a description of the agent's changes of desire and will appeal to principles about how desires do and should rationally change. I think that this is something which, though extremely basic, cannot easily be dealt with on standard accounts of rational action.[8] My description of the first of my two cases was meant to emphasise this, in an intuitive way, but the point has to be worked out in a more general setting.

Given a combination of beliefs and desires $B_1, B_2, \ldots; D_1, D_2, \ldots$, held by a single agent at a single moment of time *t*, we say that they make an action A rational, under description 'A' when some abstract relation R holds between 'A' and the set $\{B_1, \ldots, D_1, \ldots\}$. The relation may be that performing A would maximize expected utility, as given by the B_i and D_j, or that it represents the best means according to the B_i to the ends

[8] I am here following what I think of as the 'East Anglian' line on desire, see G. R. Grice *The Grounds of Moral Judgement*, Cambridge: Cambridge University Press, 1967, Ch. 1; E. J. Bond *Reason and Value*, Cambridge: Cambridge University Press, 1983; Ross Harrison 'Discounting the Future', *Proc. Arist. Soc.* 81, 1981/2; Martin Hollis 'Rational Preferences', *The Philosophical Forum*, 14, 1983, pp. 246-262.

described by the D_j, or any of a variety of other workings out of the same theme. But to say that is not to say that an agent having these beliefs and these desires will if rational perform A. For the agent may just as reasonably cease to believe some of the B_i or cease to want some of the D_j. Presumably if the fact that R holds between $\{B_1, \ldots, D_1, \ldots\}$ and A is to *explain* A, it must correspond to some process whereby the agent registers the connection between A and what he wants and thinks (not necessarily by forming a belief to the effect that R holds between them) and is thereby led to act. But if the act is later than the moment *t*—as it must be if it is more than an instantaneous act of will—then this process is something which even in the most rational of creatures may not occur. For if we take rationality just to consist in performing actions which bear R to one's beliefs and desires at the time of action then by the time of action the beliefs and desires may have changed so that they no longer bear R to A. And if we take a more subtle—and more realistic—construal of rationality, which makes an action rational if it bears some such relation to beliefs and desires the agent may rationally have at the time of action, then we can no longer be assured that A is so qualified.

VI

Equilibrium/dynamic I have been using a distinction between equilibrium and dynamic models of rationality. It is pretty straightforward, but I should now be explicit about it. An equilibrium model describes coherence, or consistency, or defensibility. These are equilibrium concepts because they concern what is satisfactory about an agent's states (including actions as states) at any one time. They say nothing about how the agent may or should change. Taken by themselves they are of limited interest, not just because coherence at each moment is compatible with the most arbitrary or peculiar changes from one moment to another, but also because most of the interesting questions about rationality are questions about what states one ought to *come* to have, what beliefs, desires, actions, or whatever one should add or subtract. And of course the states of just about all agents just about all of the time are far from any equilibrium.

Dynamical theories of rationality, on the other hand, say what changes of mind, or if you prefer what transitions between states, are rational. For example, for *the* example, any account of theory acceptance, or most of epistemology, is about the conditions under which one should come to believe various things.

My claim above was that psychological explanations, to the extent that they are formed around assumptions about rationality, must involve dynamic models. My reasons should be clear by now. What still needs to be made clear is what constraints this puts on the actual form of explanation by motive. It is, for example, quite possible that a pattern of explanation might be based on an equilibrium model of rationality plus some further assumptions which transform it into an account of rational change of desire and intention. It is in fact quite usual to derive dynamical models from equilibrium models. This is typically done by means of what I shall call the equilibrium-extension trick: one has a set of states in equilibrium and another state (or, more generally, a choice of states) which may or may not be added to them; one then takes it as rational to add the state in question if the new set resulting from adding it would preserve equilibrium.

Consider the most relevant special case where the equilibrium theory is the familiar expected utility model, according to which an agent's beliefs and desires and intentions to action (all identified by their content propositions) are in equilibrium when for any act *a*, *a* is intended iff there is no act *b* incompatible with *a* for which

$$\Sigma_i b(p_i | b) v(p_i \& b) \text{ is greater than } \Sigma_i b(p_i | a) v(p_i \& a).[9]$$

In other words, a rational agent's preferences between actions should have the same order as their expected utilities. Now to turn this into a dynamical model one applies, usually implicitly, the equilibrium-extension trick: one assumes that rational agents act so as to maximize expected utility according to their degrees of belief and desire *before* deliberation—they acquire all

[9] see R. C. Jeffrey *The Logic of Decision*, second edition, Chicago: University of Chicago Press, 1983. Note that the sum depends on the choice of a set of partition propositions P_i, which is not an uncontroversial matter, as Allan Gibbard and William Harper show in 'Counterfactuals and Two Kinds of Expected Utility', in *Ifs*, W. L. Harper, R. Stalnaker, S. Pearce, eds., Dordrecht: Reidel, 1978.

and only those intentions to action which when judged by their existing beliefs and desires maximize expected utility.

But the tacit assumptions of the equilibrium-extension trick are pretty clear, and the limits of their plausibility are also pretty clear, in this case. One thing that must be assumed to make the trick work is that by the time the act is performed the agent will still have the original desires. One drastic way of ensuring this is to make the agent's desires be based on some set of unchanging basic and permanent preferences. Call this the groundedness assumption. Note that what it really amounts to is an evaluation of changes of desire in terms of the equilibrium or lack of it that results from adding a desire to a fixed subset of one's previous desires. It is controversial in two ways: factually, in terms of the existence of (enough) such basic desires, and normatively, in terms of the advice to ignore everything else that one wants. It is also much stronger than needed for most actual cases, but some form of it is standardly used to smooth over this particular obstacle to getting from the equilibrium to the dynamical version.

Another thing we must assume is that the choice of a best action does not depend on facts which themselves depend on (this and) later choices of action. This one is a little harder to explain. But suppose that one is deciding whether to follow one's inclination to devote one's life to poetry, living off Arts Council grants and occasional visits to lesser American universities. One consideration might be whether one will later marry someone of a comfort-loving but dependent disposition, likely to be unhappy without comfort and security. The expected utility of the act of investing in one's art (instead of becoming a trainee accountant) depends on the likelihood of one's later making such commitments. But that depends on later decisions. What likelihood should one now attach to one's future decisions? There are two natural strategies. One is to think in terms of one's present estimate of the probability that one will make a future decision in a given way. This probability *might* be taken to be determined by the expected utility of the future acts by one's present lights, and thus to be either 0 or 1. This amounts to assuming that in making each present decision one is in effect committing oneself to a course of action which pre-settles all future decisions. Such a strategy makes great demands, to put it mildly, on one's confidence in one's future rationality, and in the stability of one's future beliefs and desires. (A paradox: these tend against one another—if you are at all rational you can expect your future beliefs and desires, and thus your future propensities to action, to be vastly different from your present ones.)

An alternative strategy would be to take all future decisions as undecided, to give all the relevant alternatives probability 1/2. (I'll ignore the usual problems about indifference principles.) But not only is this unrealistic in some situations, since you know more about your future self than *that*, it is actually at variance with the expected utility model. (Since, whether or not you trust them, you do have beliefs about your future decisions and the value of their outcomes.)

The groundedness assumption will help with this problem slightly, but far from completely. One condition that will at least partially tame the problem is the assumption that the acts with which we are concerned pay off quite quickly. That is, either their expected utility depends on facts which are relatively independent of the performance of results of later acts, or they can be grouped into bundles of actions which are mutually independent in this way. Let me call this the atomicity assumption.

VII

The variety of rationality Now I can pull some of the strands together. The equilibrium-extension trick will be plausible for actions which are immune to changes in the agent's desires and to uncertainties about future decisions. Either put simply this way or in terms of atomicity and groundedness, the most likely candidates for membership in this class are classical intentional actions, my subcategory of Acts: small-scale actions carried out deliberately in the course of larger projects. For the time-scale of such an act is too small for the agent's desires to have changed during its performance, and its imbedding in a larger project shifts the sensitivity to future decisions from the individual act to that project. Similarly, within the time-scale of such an act there will be a sub-class of desires which remain unchanged and in terms of which the changes of the others may be explained. The larger the time-scale the less plausible this will be.

In general, the larger the scale of an action, the more it is like an Accomplishment rather than an Act, the harder will it be to form an explanation of it around a model of the rational equilibrium of beliefs, desires, and intentions. That is not to say that it cannot be done, but the explanation will have to appeal to more substantial principles about the evolution of desire than are implicit in the equilibrium-extension trick. The result is a family of patterns of explanation by motive, which can be obtained by the following recipe: start with an interpretation of the coherence-conditions of beliefs and desires, e.g. utility maximization (itself occurring in two distinct forms), the minimax rule, or any of the infinitely many possible variants on them.[10] Then add suitable assumptions which turn it into a dynamical theory.

Suppose for example that one begins with the minimax model, according to which one should choose that action whose worst possible consequences are least undesirable. As a model of the way in which an agent with definite desires and beliefs should make decisions this has come to seem pretty implausible. But as a model of how to act on beliefs whose objectivity (even as estimates of probability) one doubts, or desires whose permanence or quantitative ranking one is not confident of, it has definite attractions. Because of one's uncertainties one cannot make a sensible estimate of how much one can expect from various courses of action, but at any rate one knows, in many cases, what it is that one wants most to avoid, and one can follow a course which avoids it. And this is exactly the situation very often when a present decision cannot be disentangled from future decisions. One often then knows the broad outline of what one wants, but must operate with probabilities which are uncertain not just for lack of evidence but because they concern

[10] What I call minimax and maximin are sometimes called maximin and minimax. See R. D. Luce and H. Raiffa *Games and Decisions*, New York: Wiley, 1957, Ch. 13. A large class of decision methods can be seen as variants of a single pattern as follows: in each of them an act is associated with a crucial quantity, which is then to be maximized. Given a partition, the crucial quantity of an act can be defined as a function of the distribution of utilities conditional on that act as a function of the partition. Different such functions give different decision-criteria: minimax uses a function which takes as value the minimum point of the distribution; for utility maximization its value is the mean of the distribution; a method commonly used in everyday life but apparently not much studied has a function taking as value the greatest point of the distribution.

the results of decisions one has not yet made. Then at least part of one's decision-making is likely to be made in terms of strategies, such as Minimax (if one is cautious, or Maximin if one is a gambler, or most likely something in between), which do not require firm estimates of probability.

The person in my earlier example, contemplating a career as a poor poet, might well decide in accordance with such a strategy. And then the strategy itself could be used as part of an explanation of the decision, if it was consistently used in the later decisions (whether to marry one person or another, whether to take this job or that, . . .) connected to it. (And the strategy used could change, too, but this would itself have to be explained, or mentioned as an assumption about the person's history.) The decision in question would thus be explained in terms of a strategy directed at the whole Accomplishment of which it is a part: the explanation appeals to a fact about the agent's character which is connected with the manner in which decisions are managed during such an extended performance.

There are in fact three important contrasts between equilibrium and dynamical treatments of rationality here. First, that uniquely best actions will be extremely rare in the dynamical case; one will speak not of the rational action for an agent to perform but of various courses of action possessing various kinds of reasonableness (e.g. caution). Second that in order to use a model of dynamical rationality as part of an explanation of an action one will have to appeal to, or implicitly assert, premises about the agent's character, emotions, or moods: an impetuous, headstrong, or simply foolish person cannot be supposed to be deliberating along minimax lines unless some further factors are appealed to. And, third, and most important here, the kind of equilibrium theory that best fits an action depends on the subcategory of action. Simple maximization best fits Acts, strategies that are less sensitive to the probabilities involved best fit Accomplishments.

This leaves us, I think, not so much with a conclusion as with a project. It is that of relating questions of the unity of rationality to questions of the unity of action. My claim about rationality is based on a rather unstartling idea: that there is no single relation between beliefs, desires, and actions such that actions are rational when they bear this relation to beliefs and desires of the agent. My twist on it is the suggestion that the variety of plausible candidates for models of rational decision making can be understood in terms of the variety of dynamical situations to which criteria of rationality apply: since actions are nearly always performed during an interval of time, and as parts of larger courses of action, changes in the agent's beliefs and desires nearly always have to be taken into account, and it is the variety of ways in which this can happen that invisibly affects our intuitions about decision making. This suggestion is tied to my claim about action, which is that much of what we say about action is invisibly relative to the particular subcategory of action in question: actions are intentional or rational not only relative to intentions but also as members of particular subcategories. The project is to make explicit the related subcategories of action, patterns of commonsense psychological explanation, and strategies of rational decision.

I have begun this project here, and done enough to show the appeal of one route through this maze of connected concepts. It is to start with natural-seeming models of rational equilibrium of beliefs, desires, and actions, and to trace their plausibility to underlying intuitions about the dynamical equilibrium of actions of appropriate subcategories. The appeal of this method stems partly from the solidity and extent of the literature on rational decision-making. I should admit to a suspicion, though, that while this is the natural and promising strategy the actual psychological picture is just the reverse: we have deeply ingrained dynamical models of belief, desire, and intention, acquired through our acquisition of the culture's concept of a person, and these shape the intuitions which lead both to our concepts of intentional action and to models of rational equilibrium.[11]

[11] The line of thought in this paper started in conversation with Robert Klamer, see his Ph.D thesis *Desire, Belief, and Intentional Action*, Bristol 1983. I am also grateful for help from Mark Cohen, Ronald de Sousa, and Martin Hollis.

THE VARIETY OF RATIONALITY

Adam Morton and David Holdcroft

II—David Holdcroft

Adam Morton's paper raises a number of challenging questions about the concepts of intentional action and of rationality. My response to these questions will, I am afraid, be rather piecemeal, since I find it difficult to formulate an overall view. However, I hope that what I say does some justice to the complex argument of his paper. Since Morton's own discussion of the concept of intentional action is only loosely related to his discussion of that of rationality, I shall follow him in treating them separately.

I

Intentional Action

Morton begins his discussion with a contrast between two views, one attributed to Austin, and one to Davidson. The former view, which Morton thinks now commands little support attacks

> the idea that we have a single category of things which people do, quite naturally described as 'intentional', which is at the same time the central example of actions in general and a prerequisite for the application of various important predicates. (p. 139)

However, a Davidsonian response to the case Austin cites in support of this view[1] would rebut the conclusions Austin draws on the basis of the example, by pointing out that an action can be intentional under one description but not another. Thus, though I tied the string in order to play cats' cradle, I did not tie it in order to set a trap for my relative. There is, therefore, no need, at least in this case, to doubt the unity of the concept of intentional action.

But though, as far as I can see, Morton does not question the adequacy of the Davidsonian response in this case, and in

[1] The response might also be dubbed 'Anscombeian'.

particular does not question the claim that an intentional action is so only under a description, he nevertheless has reservations about the Davidsonian position, and some sympathy with the Austinian one—though of course for different reasons than those Austin himself gave. The reservations have two sources, the deviant causal chain problem, and a claim, which Morton argues for at length, that there is a much greater variety of types of action than traditional accounts, Davidson's included, recognise.

It is, at this stage, worth pausing to ask exactly what is the deviant causal-chain problem. It arises on a reductive account of what it is to act with an intention described by Davidson as follows:

> If someone performs an action of type *A* with the intention of performing an action of type *B*, then he must have a pro attitude towards actions of type *B* (which may be expressed in the form: an action of type *B* is good (or has some other positive attribute) and a belief that in performing an action of type *A* he will be (or probably will be) performing an action of type *B* (the belief may be expressed in the obvious way). The expressions of the belief and desire entail that actions of type *A* are, or probably will be, good (or desirable, just, dutiful, etc.). The description of the action provided by the phrase substituted for '*A*' gives the description under which the desire and the belief rationalize the action. So to bring things back to our example, the desire to improve the taste of the stew and the belief that adding sage to the stew will improve its taste serve to rationalize an action described as 'adding sage to the stew'. (Davidson 1980, p. 86)

However, it is plainly not sufficient for the truth of '*X* did *A* with the intention of doing *B*' that *X* should have the relevant beliefs and pro-attitudes:

> Someone might want tasty stew and believe sage would do the trick and put in sage thinking it was parsley; or put in sage because his hand was joggled. So we must add that the agent put in the sage because of his reasons. This 'because'

> is a source of trouble; it implies, so I believe, and have argued at length, the notion of cause. But not any causal relation will do, since an agent might have attitudes and beliefs that would rationalize an action, and they might cause him to perform it, and yet because of some anomaly in the causal chain, the action would not be intentional in the expected sense, or perhaps in any sense. (Davidson 1980, p. 87)

So the deviant causal chain problem would be solved if we could either give an account of the causal relation that must hold between a person's beliefs and attitudes and the action they rationalise, for that action to be intentional, or, more radically, show that it is a mistake to suppose that such an account is necessary. Morton certainly does not seem to do the former thing; yet I do not find it easy to believe that he is arguing for the second alternative—though some remarks he makes suggest that he might be.

However, to discuss his argument further it is first necessary to consider the other source of his reservations concerning a Davidsonian postition, *viz.* his belief that there is a greater variety of actions than traditional accounts allow. He urges to begin with that 'What we need to do is to look at the ways in which actions can be classified into kinds in terms of the ways in which they can be explained' (p. 142). The adoption of this approach leads to the introduction of types of actions which he calls respectively 'accomplishments' and 'successes'. An accomplishment involves a result which was not necessarily intended by an agent, and is not the consequence of any single act of his.[2] Thus one might say that a player inspired his team by his overall performance, even though he did not aim to do so. A success, by contrast, requires a correlative intention: what is special about these cases is the fact that though the agent does something which produces the desired result, he does not believe of that which produces the result that it will produce it. Thus

[2] My account of accomplishments follows what Morton says on pp. 142 and 143. But on p. 151 a different account is given: 'What do Accomplishments, Successes, Unaccidents and Acts have in common? This much, I think: in all cases there is something that may be called an overall intention, . . .'

accomplishments differ from standard acts by the absence of an overall intention; whilst successes differ because the agent does not at some level of specificity know how to produce the result.

An important point, connected with this last one, is, Morton urges, that

> The only descriptions under which they are uncontroversially non-intentional turn on *de re* references to aspects of the production of the resulting states of affairs ('by pushing that key', 'by aiming at that sound'). (p. 145)

In other words, the lack of a certain *de re* belief precludes the agent from acting with certain connected intentions. This seems right. So too does the further claim that

> when we *say* of someone that they intend to do something, describing the intended act in terms of a referential link between the actor and some object or aspect relevant to the act, then we are usually ascribing to the actor a plan of acting with reference to those objects of reference, and by use of the referential tie. And in describing an act as intentional with respect to an intention, we usually mean that it was carried out by use of the connections involved in the referential tie. (p. 146)

Thus if I say that I intend to drive to Devon by car, then, assuming sincerity, you know not only what I want to achieve (going to Devon), but how I want to achieve it (by driving).

This leads Morton to suggest that

> we single out some of a person's states as intentions, not because they are different in content or psychological role from others, but because in so labeling them we are signaling a claim that their content specifies the information (better, information link) with reference to which the agent will act. And—tied to this—in saying that an act is intentional, in accord with an intention, we mean that it is guided by the information (link) alluded to in the relevant ascription of intention. (p. 146)

Thus, in describing an act as intentional we are describing not only the intention with which it was done, but the means by which it was done. Moreover, Morton argues, the problem of deviant causal chains does not arise on this account. This conclusion is, it is true, later qualified; but because of its intrinsic interest I would like to discuss the unqualified version first.

To begin with, it seems clear that we do not distinguish intentions from wants simply because the content of a description of the former 'specifies the information . . . with reference to which the agent will act' (p. 146). I can intend to go to France next summer without, at the moment, having any intention of so doing in a specific way. To the question 'How will you go?' I could reply, 'I haven't decided yet'. Moreover, it would seem that the description of the content of a want could be quite specific about the means to be adopted to gratify it, and yet remain the description of a want, not an intention.

It is, of course, true that descriptions of intentional actions often refer to the specific means used to perform them ('driving to France'); but they do not always do so ('going to France'). Sometimes, moreover, when they do not, they do not because we do not know what the means were; and sometimes they do not because questions of means have no application ('twitching my nose'). So it would seem that any solution to the problem of deviant causal chains based on the fact that sometimes descriptions of intentional acts include descriptions of the means used (and intended) to perform them will be of limited generality.

But whether or not I am right about this, the further question arises how inclusion in the description of an intentional act of the means intended to perform it helps with the problem of deviant causal chains. That problem is, it will be recalled, the problem of specifying the nature of the causal relations that must hold between a person's beliefs and attitudes and the action they rationalise, for that action to be intentional. Presumably, a mere description of the means employed will not help with that problem unless included in the speaker's beliefs and attitudes are ones about the means. Thus, suppose that my action is intentional under the description 'driving to Devon', and that (*i*) I believe that by driving I can get to Devon, and (*ii*) want to get to Devon by driving. But if (*i*) and (*ii*) are added to my other relevant beliefs and desires, the question surely remains: How does this set, including (*i*) and (*ii*), have to relate to my action for it to be intentional? The fact that the set contains true beliefs about effective means (including *de re* ones), and pro-attitudes to those means will not make it impossible for there to be deviant causal chains. The example from Davidson quoted by Morton makes just this point.[3] The problem posed by the example is that the release of the rope is not voluntary; but we cannot solve the problem by specifying that it must be, if our aim is to give a non-circular analysis of intentional action.

It is possible that I have misrepresented Morton, and that the specification of a specific kind of causal link is important for him primarily from the point of view of the classification of actions into generic types, i.e. accomplishments, successes etc. But the question then becomes: What bearing does this classification itself have on the deviant causal chains problem? One possibility Morton seems briefly to entertain is that since actions are classifiable into types, we can, if we wish, specify a type to cover any causal history however bizarre (p. 153). And this suggests that Morton may have been tempted, if only briefly, to adopt the second solution to the deviant causal chains problem, i.e. to argue that it is a mistake to suppose that it is necessary to specify the causal relation that must hold between beliefs, pro-attitudes and the action they rationalise, for that action to be intentional. It is perhaps in this spirit that he suggests that the man on the rope example seems puzzling to us because we have no familiar category of action in which to place it. But this seems to me not to be a very convincing suggestion. There is nothing abstruse or difficult about the case at a common sense level; most of us could readily think of similar cases in which the wish is involuntarily father to the act. Puzzlement only arises from a theoretical perspective in which we are trying to explain what it is to do an action intentionally in terms of a relation between an agent's beliefs, pro-attitudes, and his act. Moreover, it seems to me that the project of classifying actions in terms of the kinds of causal tie they embody, to which I am sympathetic, cannot proceed unless

[3] See Davidson 1980, p. 79: 'A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never *chose* to loosen his hold nor did he do it intentionally. It will not help, I think, to add that the belief and the want must combine to cause him to want to loosen his hold, for there will remain the *two* questions *how* the belief and the want caused the second want, and *how* wanting to loosen his hold caused him to loosen his hold.'

we can say what those causal ties are; and this we cannot do unless we have a solution to the problem of deviant causal chains. Indeed, from the point of view of the project the difficulty might well seem greater, for we have to specify a variety of causal ties, and say in each case what would constitute deviancy.

Finally, whilst on this topic, it is worth asking whether the examples given of successes are all that different from 'traditional' examples of intentional actions. There is, after all, in both cases, an overall intention, and in B(*ii*) the means adopted are effective—sooner or later the E flat key will be pressed, even though the agent does not know which one it is. It is true that in B(*i*) rather desperate experimentation produces the desired result, so that it is luck, or a fluke; but as Morton points out there seems to be no difficulty in saying what is intentional, what not. By contrast the category of accomplishments is more challenging, partly because of the absence of a need for an overall intention, and partly because what is accomplished is so by means of a series of actions, themselves possibly very diverse and separately motivated. This makes the problem of saying how accomplishments are related to the actions which sustain them a difficult one. Certainly, Morton seems right to distinguish these cases from ones in which the sub-acts (e.g. writing '*c*', writing '*a*', writing '*t*'), are not motivated independently of the act of which they are part (writing the word 'cat').

II

Rationality

Whilst I agree with Morton that any theory of rationality has to cope with the fact that our beliefs and desires change, I am not sure that I fully understand which jobs a dynamical theory taking account of this fact would be doing which existing theories of decision under uncertainty do not. In this connection, my reaction to the case of the mother who steals her child's Mars Bar is to say that the source of our puzzlement is surely the same whether she had known all along that it was the child's, or only discovered that it was after resolving to steal it. In either case it is difficult to understand why she should have stolen it, and our difficulty is to see what her values are.

It is tempting to argue that there would be a problem in the case in which she discovers that the Mars Bar is her child's after resolving to steal it, if that resulted in a change in her resolve. But if to begin with she would not have stolen it had she known that it was her daughter's, it seems to me that the change in her resolve is not very puzzling, because it is not indicative of any important change in her views of what is desirable (i.e. of her basic preferences). This is, of course, not to deny that changes in her non-basic preferences and beliefs have to be taken into account to explain why she is not now prepared to do something she was prepared to do, and that a theory about what is involved in such changes will be very complicated. But it is to say that the change in belief attendant on the discovery that the Mars Bar is her child's does not create any special perplexity for her; provided she does not fail to take her basic preferences into account, and reasons properly, she will not take the Mars Bar. The point is that changes in beliefs and desires *per se* do not necessarily create difficult problems of decision for agents. Numerous changes in either are compatible with unchanging basic preferences. The newly discovered belief that I shall inherit money next week may lead to a greatly diminished desire to work, because I attached only an instrumental value to work as a means to a standard of living, which is now secure. But my basic preference for a high standard of living remains unchanged.[4] In such cases, what Morton calls the equilibrium extension trick seems to work; for the change in my beliefs leaves my basic preferences undisturbed. What I do, assuming fixed basic preferences, is to be expected.

Assuming fixed basic preferences then, what scope would there be for a dynamical theory? Apart from having to say in what circumstances new information is acceptable, worth noting, and what should be done if it is not consistent with existing beliefs, there are many other questions about the relations between new information and existing purposes which it would have to tackle. Aaron Sloman has made a useful list:

> (*i*) Does this imply that a particular current purpose has been achieved or frustrated?
> (*ii*) Does it imply that particular current purposes are unexpectedly near to or far from being achieved?

[4] See Hahn 1982, p. 190.

> (*iii*) Does it imply that a current purpose can be achieved more efficiently or quickly or at less risk or cost, or in a more enjoyable way, etc., by modifying an on-going purpose or terminating it and starting with a new strategy: that is, is there a better way of doing what is currently being done?
> (*iv*) Does it imply that any current purposes are mutually incompatible?
> (*v*) Is this worth examining more closely to see if questions like (*i*) to (*iv*) get a positive answer after specialised investigation?[5]

The question arises then of the ways in which a theory of rationality is deficient if it is not supplemented by a dynamical theory of this sort, i.e., one which can deal with the problems just mentioned, but which assumes fixed preferences.

It is true, as Morton points out, that the fact that a person's beliefs and desires frequently change make it difficult to predict what he will do on a particular occasion, even if it is known what his beliefs and desires are, and he is a known maximiser of expected utility. For as Morton says, ignoring the possibilities of miscalculation, lack of a relevant skill, or opportunity, a person may, even when A bears the relation R to the set $\{B_1, \ldots, D_1 \ldots\}$, not, on reflection, do A because of some change in his beliefs or desires. But I am not clear that it is a ground for objection to the theory that A is rational if it bears R to the set $\{B_1, \ldots, D_1, \ldots\}$; after all, many versions of this theory do not purport to be predictive.[6]

However, equilibrium theories have also been criticised on the grounds that they are limited to taking an agent's desires and beliefs as given, however bizarre they may seem.[7] Moreover, as Morton points out, the variation in an agent's beliefs and desires from moment to moment may be very capricious; but this is a matter of indifference to an equilibrium theory. However, if we had a theory of the rational evolution of belief and desires, we

[5] Sloman 1978, p. 131.

[6] See Jeffrey 1965, p. 155.

[7] 'The Bayesian model may be as applicable to the deliberations of a knave or a fool as to those of a good and wise man, for the numerical probabilities and desirabilities are taken to be subjective in the sense that they reflect the agent's actual beliefs and preferences, irrespective of factual or moral justification.' (Jeffrey 1965, p. 1)

would perhaps be able to say what it is for a beliefor desire to be rational, namely, that it does not violate the principles of the appropriate dynamical theory. Thus, it might be proposed that in cases involving risk

> A is rational, if and only if,
> (*i*) A bears R to the set $\{B_1, \ldots, D_1, \ldots\}$
> (*ii*) No B_i or D_i violates the principles of the appropriate dynamical theory.

On this account a dynamical theory would be an essential part of a total theory of rational action, and not just an extra refinement. So that an inability to say anything about changing desires and beliefs would be a serious defect, even assuming fixed basic preferences, if this proposal were correct; though we would still not, I think have a predictive theory.

Some things Morton says (p. 157) suggest that he has these sorts of dynamical theories at least partially in mind. However, rather than speculate further about such theories, I would like, in the remainder of this paper, to say something about the kind of problems of changing beliefs and desires that trouble Morton most.

These cases are ones in which basic preferences cannot be assumed to be permanently fixed. Since they may change, my present attempts to act in conformity with my basic preferences may seriously inhibit my chances of satisfying ones I will later come to have; or, as Morton says, it may be reasonable for me to embark on a particular course of action now, only if I can be sure of something very unreasonable, namely, that my basic preferences will not change. Morton has given an example of the second difficulty; here is one of the first:

> Jane is going to University. She is passionately interested in philosophy, and believes devoutly in communal living, vegetarianism and nuclear disarmament. However, she knows that many people who once were students and held views like these end up despising philosophy, like to make money, live a comfortable married existence and eat meat. She also sees that if she lives her life to the full now, she may find it difficult to do what she most wants later. What then should she study at university?

Morton thinks she might adopt a minimax strategy, for instance read Management Science and Philosophy, rather than straight Philosophy. Nagel, commenting on a similar case, argues that there is no problem in her doing this provided that she thinks of the changes in her basic preferences as only changes in preferences.[8] For she could then have second order preferences about her first order basic preferences, e.g. to do that which enables her to fulfil as many first order basic preferences as possible at each stage of her life. Certainly, the balancing of present against future gratification required by following this policy is not incoherent, if only preferences are involved. Moreover, one presumably has to decide on the basis of the preferences and values one has now, and the suggested second order principle about first order preferences is quite plausible from the point of view of self interest. However, if a minimax policy has some plausibility in this case, then whilst it is true that it takes into consideration problems posed by the dynamics of belief and desire, it is not clear to me that we are dealing with a novel kind of 'dynamical' theory. In other words, this seems to be just a special case of the application of the theory of decision under uncertainty.

However, this solution does not enable Jane to adopt a policy that will cope with all changes in her preferences. This becomes clear when we ask why should Jane treat her second order preferences as any more immutable than her first order ones? If they are only preferences, why should they too not change? And if Jane envisages that they might change, what should she do? We seem to be back with just a more complicated version of our original problem. For if we cannot suppose that our present first order basic preferences have any special status because they are present, we surely cannot suppose that our second order ones have a special status for that reason. So if Jane has adopted a strategy of tolerance it may well seem to her later that she should not have.

But from the point of view of psychological explanation how plausible would it be to suppose that Jane has adopted such a policy? I think that even supposing that we are dealing only with preferences, it may not seem very plausible. For to adopt it she

[8] Nagel 1970, p. 74.

would have to be prepared to think it reasonable to have preferences quite different from her own present ones. And if she is unable to imagine what it would be like to be someone who has basic preferences radically different from her own, then a bias towards the present would seem inevitable. Moreover, the number of ways in which her basic preferences might change may be so great that she might think it pointless here and now to guard against all contingencies, and best simply to do what she now most wants. Even so, attribution of the policy of tolerance to her may have some plausibility.

However, as Nagel argues, if what we have been treating as her basic preferences are not just preferences, but include moral values and principles, then the case is completely altered.[9] For although she can indeed envisage that these may change, she can hardly trade off her present ones for the ones she may have in the future which are bound now to seem repugnant to her. This certainly seems correct. Admittedly, it might be suggested that Jane might have as one of her moral principles the principle that people ought to act on their moral principles; so that we can take the same line in this case as we took in the previous one which involved only preferences. However, I find the suggested principle unappealing; surely, though we may respect someone who acts on principle, that does not affect our view of the worth of his action, if we think what he has done is wrong.

I am not sure what conclusions about Morton's project of psychological explanation should be drawn from this. I agree with him that any adequate theory must deal with the decisions people take to deal with uncertainty, and that this involves uncertainty about their own future beliefs and desires. I agree with him too that there are specially difficult problems when the changes contemplated involve basic preferences, and that in neither case can we understand what people are doing unless we can understand the policies they have adopted to deal with uncertainty. But while I agree that the nature of some of these policies can be illuminated by the theory of decision under uncertainty, I find it difficult to see that the kind of case in which someone recognises that his moral values may change is one that would involve policies of this sort.

[9] *Ibid.*

REFERENCES

Davidson, D., 1980, *Actions and Events*, Clarendon Press.
Jeffrey, R. C., 1965, *The Logic of Decision*, McGraw Hill.
Hahn, F., 1982, 'On Some Difficulties of the Utilitarian Economist', in eds. A. Sen & B. Williams, *Utilitarianism and Beyond*, C.U.P.
Nagel, T., 1970, *The Possibility of Altruism*, Clarendon Press.
Sloman, A., 1978, *The Computer Revolution in Philosophy*, Harvester.