

Physicalism and the Bogey of Determinism

[Aaron Sloman](#)

[School of Computer Science](#)

The University of Birmingham

(Written when I was at the University of Sussex).

Presented at an interdisciplinary conference on Philosophy of Psychology at the University of Kent in 1971. Published in the proceedings, as

A. Sloman, 'Physicalism and the Bogey of Determinism'
(along with Reply by G. Mandler and W. Kessen, and additional comments by Alan R. White, Philippa Foot and others, and replies to criticisms)
in *Philosophy of Psychology*, Ed S.C.Brown, London: Macmillan, 1974, pages 293--304. (Published by Barnes & Noble in USA.)
Commentary and discussion followed on [pages 305--348](#).

This paper rehearses some relatively old arguments about how any coherent notion of free will is not only compatible with but depends on determinism. However the mind-brain identity theory is attacked on the grounds that what makes a physical event an intended action A is that the agent *interprets* the physical phenomena as doing A. The paper should have referred to the monograph [Intention \(1957\)](#) by Elizabeth Anscombe (summarised [here by Jeff Speaks](#)), which discusses in detail the fact that the same physical event can have multiple (true) descriptions, using different ontologies.

My point is partly analogous to [Dennett's](#) appeal to the 'intentional stance', though that involves an external observer attributing rationality along with beliefs and desires to the agent. I am adopting the *design* stance not the intentional stance, for I do not assume rationality in agents with semantic competence (e.g. insects), and I begin an attempt to explain how an agent has to be designed in order to perform intentional actions; the design must allow the agent to interpret physical events (including events in its brain) in a way that is not just perceiving their physical properties.

Some of ideas that are in the paper and in my responses to commentators were also presented in [The Computer Revolution in Philosophy](#) (1978), including a version of [this diagram](#) (originally pages 344-345, in the discussion section below), discussed in more detail in [Chapter 6](#) of the book, and later elaborated as an architectural theory assuming concurrent reactive, deliberative and metamanagement processes, e.g. as explained in this 1999 paper [Architecture-Based Conceptions of Mind](#), and later papers, and crudely depicted [here](#).

The original paper follows, preserving page divisions. This file should work on all browsers, having passed two validation tests at [www.w3.org](#). A PDF version based on the scan, without the discussions, which may print in a better format, is available as

<http://www.cs.bham.ac.uk/research/cogaff/sloman-bogey.pdf>. Another, more complete, PDF version derived from the html version is [here](#).

(I may later add further notes and comments to this HTML version.)



(Last updated: 29 Dec 2005)

Physicalism and the Bogey of Determinism¹

AARON SLOMAN

1. THE PROBLEM AND THE PROGRAMME

Only a dreadful prig could seriously condemn secret lust as a form of adultery. However, even someone who claims to be wholly concerned with the things of the mind must have some interest in what he says and does, and these require the occurrence of bodily processes. Realising that so much of what matters to us involves physical events and processes, it is natural to find alarming the suggestion that all physical behaviour of our bodies can be explained in terms of the 'mindless' workings of laws of nature. Consequently, many philosophers have tried to prove it isn't so.

More precisely, the attempt to refute physicalism (the theory that human bodies are physical systems) may be motivated by the assumption that it implies all the following: that all our actions are predictable; that mental phenomena are identical with or composed of physical phenomena; that all human actions have purely physical explanations; and that there is no room for beliefs, desires and other mental phenomena to influence our actions. I shall argue first of all that since the question whether human bodies are physical systems is an empirical one, attempts at philosophical refutation are futile, and secondly that physicalism does not have the implications just mentioned. This will require a

¹I am indebted to several colleagues, including M. Boden, D. Booth, J. W. Burgess, M. B. Clowes, J. Dorling, M. Ireland, R. Poole, T. L. Sprigge, A. R. White, for comments on an earlier version of this paper, and to N. S. Sutherland, many of whose arguments I have borrowed from his 'Is the brain a physical system?', in R. Borger and F. Cioffi (Eds.), *Explanation in the Behavioural Sciences* (Cambridge University Press, 1970) pp. 97-122.

lengthy discussion of the relation between actions and movements, or between mind and body.

First, let us be clear which thesis is under discussion.

2. PHYSICALISM DEFINED - IT CANNOT BE REFUTED PHILOSOPHICALLY

To avoid semantic problems arising out of the future growth of physics, let us agree now to use the word 'physics' to cover what is called physics in present day physics departments in respectable universities. This embraces a set of empirical and theoretical concepts, including a set of scales of measurement, a body of theory and a host of experimental and mathematical techniques. The existence of borderline cases arising out of current disputes at the research frontiers makes no difference to our present concerns. (I have not included control engineering as part of physics, though it is relevant to much of the discussion, since its concepts are intermediate between physical and psychological concepts.)

Whatever may be unclear about what falls within physics thus defined, it is indisputable that physics does not include such concepts as 'smile', 'want', 'jealous', 'reply', and that the theories of physics do not include such statements as 'If a person prefers X to Y, but knowingly chooses Y rather than X, then he must have some other preference, hope, fear, dislike or attitude to which X and Y are relevant'. It is equally clear that human bodies contain physical substances, that they have many physical properties, that many physical processes occur in them, and that physical movements and other physical processes are involved whenever we say or do anything.

Using the term 'physical behaviour' to describe any movement, process or change of state *completely* characterisable by means of concepts of physics, we can sum up by saying that it is an obvious empirical fact that many of our actions in some sense *involve* physical behaviour. One of the problems to be discussed below is what sort of involvement this is.

Another problem is whether the following 'Physicalist' thesis is true, and what its implications are:

(P) *All the physical behaviour (in the sense just defined) of human bodies or parts thereof conforms to current physical theories.*

By 'conforming to' current physical theories, I mean not merely failing to refute such theories, but also being predictable and explainable on the basis of such theories, except in so far as the theories themselves require certain physical happenings (e.g. some subatomic events) to be unexplainable. Since current physics is compatible with the existence of a mechanism in which there is significant coupling between random subatomic processes and large scale bodily movements (e.g. a robot whose changes of speed and direction of motion are partially controlled by switches triggered off by radiation from a lump of uranium inside it), thesis (P) does not imply that *all* the physical behaviour of human bodies is predictable on the basis of current physical theory and suitable measurements made in advance. However, it is clear that not much of our behaviour is significantly coupled with random processes, since if it were humans would be much less reliable and predictable than they are.

(P) is obviously a special case of the more general thesis:

(P¹) *All physical behaviour, (i.e. of human bodies and everything else) conforms to current physical theories.*

Both (P) and (P¹) are empirical. Whether the physical behaviour of some object conforms to current physics is clearly an empirical question, even though it may be very difficult to discover the answer. It is relatively easy to establish beyond all reasonable doubt by examination of its mechanism, that the physical behaviour of a clock conforms to current theories, even though there is always the possibility that more refined measurement or observation may show that the examination missed something such as an incongruous relation between the molecular structure of the spring and its tension. It is not easy to establish beyond all reasonable doubt that the behaviour of some system does *not* conform to current physics, least of all when the system is as complex as the human brain. Still, it is conceivable that the

behaviour of a relatively simple part of the brain may one day be shown, beyond all reasonable doubt, to refute some generalisation of modern physics, especially if the behaviour is then adequately explained by some quite new theory. I have gone into some obvious details here because it seems that some philosophers who argue against determinism fail to realise that if successful, their arguments would refute an empirical theory.

(P¹) is a version of the thesis of universal determinism, perhaps the only sort of version worth taking seriously in view of the indeterministic implications of modern physics. (P) is a more modest version of the same thesis, restricted to the human body. Philosophers have occasionally tried to show that determinism is incoherent or internally inconsistent, but such attempts are futile as far as (P) and (P¹) are concerned, since they are empirical and therefore not amenable to philosophical refutation. This does not rule out the possibility of an empirical refutation based on common-sense knowledge about what humans can do, but this is not as simple a matter as might be thought since (a) it would require a much clearer grasp of the totality of possible physical mechanisms than anybody has at present (e.g. it is not at all obvious that every physical information-processing mechanism must be a Turing machine), and (b) it would require an account of the relation between common-sense concepts used in the description and explanation of human behaviour and the concepts and measurement scales of physics. It is worth dwelling on (b) in order to get a clearer understanding of what (P) does and does not imply.

Normal perception and thought about our own behaviour and the behaviour of others use concepts and categories which are not part of physics. When we say (or think) such things as 'W has at last made up his mind which job to accept', 'X beat up his wife in a fit of temper', 'Y walked hurriedly towards the door', 'Z found a way of refuting Fermat's last theorem yesterday', our descriptions and explanations use concepts which presuppose that we are referring to conscious agents with beliefs and intentions concerning the actions they perform. Since common-sense knowledge about human behaviour uses such concepts, and

since (P) does not explicitly say anything about *human* actions or *mental* phenomena, there is no direct conflict between (P) and common sense. However, our problem is whether *indirect* conflict is possible, and if so how. The answer must depend on the relation between human actions or mental processes and the underlying physical behaviour.

3. THE RELATION BETWEEN PHYSICAL AND PSYCHOLOGICAL LEVELS OF DESCRIPTION

What is this relation? It is by now a commonplace that action descriptions and physical descriptions operate at different 'levels'. (See Waismann's article on 'Language strata' in *Logic and Language*, Vol. 2, ed. A. G. N. Flew.) But exactly what this distinction of levels amounts to, and what the relationship is between the different levels remains to be made clear. Unfortunately the relationship is very complex: there is not just one difference between the level of physical description and the common-sense level, but several. For instance, one difference is that the use of certain common-sense descriptions and explanations expresses moral or other values of the speaker ('As soon as the alarm sounded he made a very cowardly dash for the exit'), or may express a preparedness in principle to treat the subject of the descriptions in a certain sort of way, such as a preparedness to feel pity should the subject suffer, or a willingness to praise or blame the subject for what he has done. Such connotations about the speaker's values or attitudes are missing from descriptions in the terminology of physics. But the particular relation between the level of physics and the level of actions and mental phenomena that I want to discuss is that of *interpretation*.

It is tempting to suppose that when the performance of an action involves certain physical behaviour (e.g. certain movements, or the production of a pattern of acoustic radiation from the mouth) the action is somehow *composed of* the physical behaviour, or *constituted by* it. A stronger claim is that the two are *identical*. That both claims are false follows from the fact that the physical behaviour has to be *interpreted* (by the agent and by others) as the action, and different interpretations are possible, depending on the immediate social or psychological context. This is most

obvious in the case of speech, for the utterance of the same sequence of sounds may, on different occasions, be interpreted as the making of quite different statements. But even the physical behaviour involved in my action of walking to the door may on another occasion be involved not in walking to the door but in doing the exercises recommended by a physiotherapist. A movement of the hand may in one culture be interpreted as a friendly gesture, in another as threatening.

Thus, what action is performed when a certain physical behaviour occurs is relative to a *mode of interpretation*. When I observe and describe someone's action I normally interpret the physical configuration which I perceive, in the light of an enormous amount of knowledge which I normally use automatically. I may use knowledge of the normal mode of interpretation of that kind of behaviour in my society. I may use knowledge of the mode of interpretation most likely to be employed by that particular agent (i.e., what he intends himself to be doing, and also how he interprets what he observes his body to be doing: these can, but don't normally, come apart), and this may be based on a deep personal knowledge of the agent, or merely on some arbitrary agreement we have entered into, or on what he has been doing or saying in some short preceding interval. I may use knowledge of the normal, and generally recognised, *function or purpose* of some object or equipment he is manipulating, such as a pencil, or light-switch, and interpret him as using it for its normal purpose. Some aspects of my interpretation of the physical input or my sense-organs may be largely (though usually not entirely) determined by my genetic make-up, such as the automatic peripheral processing done by the optical system which gives me information about colours, edges, textures, movements, orientations, etc., of the objects I see, or the decomposition of complex sound waves into distinct sounds each with its own pitch and timbre. The normal adult human being has enormous interpretative resources: his visual apparatus can quickly find a suitable syntax (or 'grammar') for analysing the structure of a wide variety of visual configurations, and interpretative rules for interpreting such structures as objects standing in certain relations, with functions of certain sorts, or as persons

consciously performing certain actions, or as a social situation involving several persons with complex interactions between them (e.g. observing a group in the middle of a heated argument). That we find this so natural, that we do it all effortlessly, and that we normally all agree in our interpretations, obscures the fact that interpretation is involved. (For similar comments on the perception of pictures see E. H. Gombrich *Art and Illusion*, and N. Goodman *Languages of Art*. Max Clowes helped me to understand the importance of interpretation.)

But if *being composed of* (like being identical with) is a two-termed relation, then it cannot be the relation between physical behaviour and the action it expresses (or represents), for the latter relation is relative to a *mode of interpretation*, M, which can vary. Thus there is a three-termed relation between the physical phenomenon X, the action (or psychological phenomenon) Y, and M. Which mode of interpretation is the correct one will, as I have indicated, depend on different factors on different occasions. Sometimes there are different ways of interpreting the behaviour, none more correct than the others, in which case the behaviour is ambiguous, perhaps intentionally so. Sometimes no interpretation can be found: it is a 'meaningless' twitch or spasm, or whatever. People who find it unusually difficult to interpret the behaviour of others, or whose behaviour most normal people cannot interpret, may be regarded as mentally ill. (This issue is complicated by the fact that there are so often several layers of interpretation.) That what a physical or geometrical configuration expresses or represents is relative to a mode of interpretation is part of the explanation of the fact that there are no *geometrical* theorems to the effect that certain configurations of lines represent (or look like) faces, and also the fact that it is not possible to give necessary and sufficient conditions in *purely* physical and geometrical terms for something's being a picture of a smiling face, or for something's actually being an act of smiling at another person.

Since the relation between action descriptions and descriptions of physical behaviour involves a third term which can vary, there are very few direct implications between the two

levels of description. This is again particularly clear in the case of linguistic actions: 'He asked me the time' reports an *action* unambiguously, but leaves quite unspecified what the physical vehicle was: whether a spoken sentence was uttered, and if so in which language, and if not whether the request was written down, or perhaps made in some sign-language, etc. Even when the relevant mode of interpretation and the form of physical behaviour are partly specified, as in 'He asked me, out loud, in English, what the time was', the possible range of physical configurations compatible with the action description remains very large, as may be seen by comparing harmonic analyses of tape recordings of different people saying 'What is the time?'

Now one of the things which is not part of common-sense knowledge, at least not explicitly, is exactly what the various modes of interpretation are which we use so frequently. The difficulty of making them explicit is illustrated by the difficulty of formulating a theory of the semantics of a natural language: this is difficult even for native speakers. But without explicit formulations of our interpretative resources, which would probably require the combined efforts of psychologists, linguists, art historians, computer scientists and philosophers, it is not possible to spell out in detail the physical implications of common-sense knowledge and beliefs about what humans can do. Thus, any empirical refutation of thesis (P) *on the basis of the indirect implications of common-sense knowledge*, seems to be out of the question at present.

4. THE SUPPOSED ALARMING IMPLICATIONS OF (P)

The desire to refute one or other form of determinism seems to be very strong in many people: this is the only way in which I can account for the proliferation of very bad arguments against it which I have read and heard in philosophical discussions. Apart from irrational and emotional factors which may explain such a desire, I believe there are also some straightforward mistakes as to what is implied by a thesis such as (P). For instance, it may be mistakenly thought that (P) implies that there are no such things as mental states or processes, or that all our actions can be explained (at least in principle) purely on the basis of current

physical theory, or that even if there are mental phenomena they can have no influence on our actions.

However, a careful reading will show that neither (P) nor (P¹) says anything about the existence or non-existence of non-physical things, or mental phenomena in particular. At most (P) and (P¹) imply that all physical events and processes can be explained without reference to any such things. (Not even within the margins of quantum indeterminacy can 'non-physical' causes influence physical events, since that would interfere with physically predictable probability distributions.) Thus (P) is compatible with epiphenomenalism.

Does (P) imply that there are physical explanations of all our actions? If my action A involves physical behaviour B, and there is a complete physical explanation of B, does this not thereby explain the occurrence of A? No, for suppose S₁ is a statement describing B (and therefore using only concepts of current physics) and S₂ a statement describing the action A. Then in agreeing that this occurrence of B is to be interpreted as the performance of A, we are using some mode of interpretation, call it M. Suppose E₁ is the statement explaining, in terms of current physics, the behaviour described in S₁. Now, I think that without committing ourselves to any detailed analysis of the concept of explanation we can say that there must be *some* relation between E₁ and S₁ in virtue of which the former provides an explanation of the behaviour. For instance, it may be that E₁ logically entails S₁ or logically entails that what S₁ says is probably true. But because S₁ does not entail S₂, or stand in any two-termed relation analogous to entailment (recall that A is not composed of B), the only way to get an explanation of the action out of E₁ is to conjoin with it a statement to the effect that B is to be interpreted according to M, for according to a different mode of interpretation S₁ may be true and S₂ false. It may be that this enlarged statement stands in the required relation to S₂. But then even if it does give an explanation of the action, it will not be a purely physical explanation for such statements about modes of interpretation are clearly not parts of physics. Thus (P) does not entail that there is a purely physical explanation of every human action.

It might be thought that (P) entails that there cannot be psychological explanations of our actions, for instance in terms of the agent's desires and beliefs, because if there were such an explanation of an action it would also explain the physical behaviour involved in the performance of that action, thus invalidating the supposed physical explanation. But if E_2 is such an explanation of the action described in

S_2 , the same argument as before shows that on account of the elasticity of the relation between S_2 and S_1 , E_2 need not provide an explanation of the behaviour described in S_1 , even though S_2 gives an interpretation of that behaviour. Indeed,

one can know what action a person performed, and can have a perfectly adequate psychological explanation of it, without having any idea what sort of physical behaviour was involved, as the example of asking the time shows. This is like knowing a computer's solution to a problem, and being able to explain its solving the problem by specifying the methods the computer uses, without being able to describe or explain the physical output of the computer (e.g. a complex wave-form on magnetic tape). One may not even know whether the computer uses valves, transistors, or components of a quite different kind. But there must be *some* underlying physical behaviour with its own explanation. (This point is discussed more fully in Margaret Boden's *Purposive Explanation in Psychology* (Harvard University Press, 1972).)

A similar problem can be raised in terms of the concept of 'cause' or 'influence', instead of 'explanation'. It might be thought that (P) implies that such things as beliefs and desires cannot influence our actions, since it implies that the physical behaviour of our bodies is fully determined by physical antecedents and processes, leaving no room for psychological phenomena to make any difference. If beliefs and desires can influence our actions, then since actions involve physical behaviour of our bodies, it follows that beliefs and desires can influence physical behaviour of our bodies. How can non-physical things like beliefs and desires *influence* physical events and processes if the latter are completely determined physically? It seems that either (P) is false or mental phenomena do not influence our actions, or else influencing actions is possible without influencing the

underlying physical behaviour. The concepts of 'cause' and 'influence' are treacherous and it is worth reflecting on this last suggestion that the relation between our actions and the physical behaviour of our bodies does not include any causal connections.

It would be possible to argue that there are no causal connections between physical phenomena and either actions or mental phenomena, if, as has often been supposed since Hume, the existence of a causal connection were nothing more than the existence of a predictively reliable inductive correlation between two general types of events, or properties or states of affairs. For if the relation between physical phenomena and actions or mental phenomena is not a two-termed one, but involves also a third term, a mode of interpretation, then since different modes of interpretation may be appropriate in different circumstances, there need be no reliable inductive correlations between physical phenomena and actions or mental phenomena. Thus, it does not follow from (P) that empirical investigation could yield such correlations, or that instruments recording physical processes in our brains can be used, on the basis of such correlations, to give information about the contents of our minds.

There are indeed many good correlations between physical and mental phenomena, but these concern only *general* feelings or moods, or *general* aspects of cognitive functioning. It may be possible to tell from the physical or chemical processes in a person's body that he feels fear, or anxiety, or depression, but it is unlikely that generally applicable procedures could reveal that he fears that an attempt will be made to blow up the aeroplane in which he is travelling, or that he is anxious about his prospects as a candidate in the next election, or that he is depressed about his daughter's failure to win the local beauty competition. Electronic instruments may be able to register whether I am conscious or not, but not that I am conscious that the ticking of the clock in the hall is slower and less regular than usual. If a particular part of the brain is solely concerned with decision-making, there may be a neuro-physiological indication that I am taking a decision, but not that I am deciding whether to drive to London by car, despite the greater expense, or to go

by train, despite the greater inconvenience. For it is likely that the *languages*, or *codes*, used by human brains vary not only from culture to culture (think of the occurrence of the above moods, feelings, thoughts, etc. in an Englishman, a Frenchman and a Chinaman), but also from one individual to another, since the sense which I associate with a proper name or other referring expression, and to a lesser extent with descriptive expressions too, depends to a considerable extent on exactly what I know about the thing (or things) referred to, that is, it will depend on my previous learning theory. (I have argued this more fully elsewhere.) This linguistic variation is limited by social constraints as far as our written and spoken language are concerned, for we could not otherwise communicate with one another. But there are no such constraints on the modes of representation used by our brains for the storage and processing of the enormous amount of information each of us has to handle: so tremendous individual variation is possible.

The point is simply that what makes some particular physical configuration or process in my brain (or physical output of my body) have the function it does, or express or represent what it does, depends on a complex set of interrelationships with other things in the brain and their relationships (via sense organs and various kinds of motor output systems) to a complex set of present and past phenomena, including cultural phenomena, outside my body. An extreme case should make this clear. When a Chinaman and I both hear a spoken Chinese sentence there may be a very similar pattern of electrical activity produced thereby in certain parts of our brains. But I can interpret the sound only as *somebody talking*, whereas the Chinaman will perhaps hear a detailed report of a horrifying disaster in which his children have died.

This shows that (P) is perfectly consistent with the view that it is a complete waste of time looking for *general* (i.e. interpersonal) inductive correlations between mental contents and brain phenomena. Neither does acceptance of (P) commit one to taking seriously the prospect of fiendish neuro-physiologists using instruments to read the contents of our thoughts, beliefs, intentions, etc., by direct physical manipulation of our brains. Of course, if the physical

behaviour of our brains conforms completely to current physical theories, then that does imply that it is possible for a physical system to interpret physical phenomena in terms of psychological categories, for that, in effect, is what a brain does. But the preceding discussion shows that (P) is perfectly compatible with different brains doing things in different ways. So the problem of designing an instrument which will read my mind off my brain may be no less than the problem of designing a replica of *my* brain (or parts of it). The output of such an instrument will then be related to the contents of my brain in much the same way as my verbal output is related to the contents of my brain.

Despite the interest and importance of this point that physical phenomena are related to actions and mental phenomena via variable modes of interpretation, and therefore need not be significantly correlated with one another, it does not prove that there are no causal connections between the two levels, for the argument rests on an analysis of the concept of 'cause' which is at least controversial. In any case, it is an indisputable fact that by giving information I can change someone's beliefs, which can cause a change in his desires and intentions and thereby influence his actions *and* the physical movements of his body. Similarly, the physical process consisting of the arrival at his ear-drums of a certain pattern of acoustic radiation *can* cause a change in his beliefs, desires, intentions and actions.

For, in such cases, the mode of interpretation is fixed. *Given* that the physical phenomena are interpreted according to mode M, then certain changes at the physical level are necessary for changes at the level of actions, intentions, etc., and conversely certain changes at the latter level necessarily involve certain changes at physical level. Admittedly, since M can vary, these changes at different levels are not *in themselves* necessary and sufficient conditions for one another. But in a situation where the mode of interpretation is fixed (e.g. by social conventions and the brain structure of the persons concerned), occurrences at one level can be necessary, or sufficient for occurrences at the other.

So, despite the logical independence of events and processes at the two levels, and despite the lack of any *general* inductive correlation between them, nevertheless in a context

in which the mode of interpretation linking the two levels remains fixed the two levels are inextricably intertwined, and processes at one level are only possible if certain processes occur at the other level. So, when it is argued that if (P) is true, then our physical behaviour, and therefore also our actions, are fully determined by antecedent physical events and processes, leaving no room for such things as beliefs and desires to influence our actions, the mistake is to apply the mode of interpretation which links our physical behaviour with our actions, while forgetting that the antecedent physical events and processes in the brain are also to be interpreted in terms of psychological concepts. To argue that the beliefs, desires, etc. have no influence because changing the physical causes must change the actions, and changing the beliefs, desires, etc. will not influence the actions so long as the physical antecedents are not changed, is to forget that so long as the relevant modes of interpretation remain fixed the physical antecedents cannot be different unless the psychological antecedents are, and the psychological antecedents cannot be different unless the physical ones are.

Of course, it is possible to interfere with the physical workings of my brain in such a way that the original interpretative system is no longer operative, for instance by destroying part of my brain or giving it excessive electrical stimulation: but then the resulting physical behaviour of the body is no longer interpretable at the level of actions. It may, for instance, be something like an epileptic fit. That physical occurrences can totally disrupt mental life and the performance of actions, is thus not a new paradoxical consequence of (P), but a familiar fact which any theory of the mind—body relation must accommodate.

5. THE INCOMPLETENESS OF THIS DISCUSSION

I am painfully aware that I have barely begun the difficult task of describing the complex and confusing relation between the different levels of description and explanation. We seem to need new organising concepts here to bring out clearly the difference between the case where two logically independent processes can interact because one is an interpretation of the other, and the case of two separate processes at

the same level causally interacting because they are linked by wires, levers, radiation, etc. The latter case is a misleading analogy which keeps intruding and generates apparent paradoxes. In a mechanism like a computer or a brain, there is a kind of organisation which ensures that, so long as the system is working normally, the physical structure *constrains* the physical processes that can occur, in such a way that they are all, so to speak, 'harnessed in the service of processes at a different level. But the relation is not like that between an engine and the pump or other machine which it drives, but like that between the configuration of charcoal on the surface of a sheet of paper and the picture of a smiling face, which we take it to be. The relation is at once remote, since it is mediated by a mode of interpretation, and intimate, since, given the mode of interpretation, the existence of either the physical object or the picture is necessary and sufficient for the existence of the other. It is this intimacy which has misled some philosophers into thinking the philosophically important relation between the two levels is one of identity, or composition.

I have tried to show that the attempt to draw alarming or paradoxical conclusions from the thesis (P) may rest on a failure to grasp the complexity of the relation between physical states and processes on the one hand and our actions, beliefs, desires, decisions, etc., on the other. No doubt I have generated more problems than I have solved, for I have said little about the contents of the modes of interpretation, nor explained in detail how to tell which mode of interpretation is correct in any particular case. Neither have I shown what sort of organisation of a physical system can constrain its behaviour in such a way as to make it interpretable in terms of psychological concepts. These are matters for further inquiry. But the existence of computers which everybody agrees to conform to thesis (P) even though they can calculate, solve problems, search for significant correlations in a mass of data, control complex machines, etc., shows that concepts operating at different levels can get a grip on the same bit of the world.

6. CONDITIONS SUFFICIENT FOR THE APPLICABILITY OF PSYCHOLOGICAL CONCEPTS

As a step towards showing how psychological concepts can get a grip on a system whose physical behaviour conforms to physical theory, I shall now attempt to list a set of conditions whose conjunction seems to be a sufficient condition for the applicability of concepts like 'believe', 'want', 'decide', 'intend', and verbs of action, as we normally understand them. There are immense difficulties in practice in designing a physical system which satisfies these conditions (despite the optimism of the edition of *Science Journal* entitled 'Machines Like Men', Vol. 4, No. 10, October 1968), but there does not seem to be any logical or conceptual impossibility. Each condition presupposes the satisfaction of some or all of the preceding conditions, and adds a new type of sophistication. My formulations here are sketchy: greater precision would require too much space. The conditions follow.

1. The system is an integrated whole whose parts are controlled by or provide input to a central processing machine. It contains *sensors* capable of receiving stimuli of various sorts from the environment and motors (e.g. limbs and muscles capable of changing the environment, changing its relation to the environment (e.g. position and orientation) and possibly changing the positions and orientations of its own parts relative to each other.

2. Its input processors are capable of analysing and recognising at least some of the sensory patterns, including temporal patterns, that humans can perceive, such as two- and three-dimensional shapes and their movements, and perhaps certain sound patterns. The processing is normally quick enough for configurations to be recognised before they change.

3. It can organise at least some of its output into patterns that humans can recognise and describe such as picking something up, moving three yards, and putting it down, drawing simple polygons, or producing recognisable sound-patterns. These types of output are among the configurations its input processors can recognise.

4. The central mechanism can use the input from the sensors, possibly together with feedback from its output, as a

basis for constructing an internal model or representation of some aspects of the environment and its own relations to the contents of the environment (e.g. its position and direction of motion).

5. It is capable of *using* this representation as a basis for moving around the environment and as a basis for constructing hypothetical, or provisional, representations of previously unknown aspects of the environment, these representations then being modified, discarded, retained as provisional, or built into the main representation of the environment, depending on new input.

6. It has a set of long-term goals (including the goal of constructing as complete and accurate a representation of the environment as possible, and maintaining itself in good working order), and a set of transient goals, and some or all of its behaviour is directed to the attainment of these goals. The goals are to some extent ordered as to priority, and there is a mechanism for imposing further ordering or generating new goals (either randomly, or in relation to higher-order goals) when conflicts arise between two or more goals.

7. It is capable of using its representation of the environment as a basis for working out possible or efficient means to the achievement of its goals (e.g. working out a route to get to a place or type of object required by one of its goals), in which case it adds the adoption of those means to its store of goals (subject to the above condition about conflicts).

8. It can construct a representation of some aspects of its own internal states and processes, including such things as what its goals are, what the form and contents of its representation of the environment are, what procedures it uses to select means or resolve conflicts between its goals, etc. It can change its internal state when this is a means to the achievement of one of its goals.

Of any system satisfying all these conditions it seems to me that it would be perfectly in order to use the following descriptions as I understand them: 'There are some things it knows, and some things it believes, though mistakenly.' 'There are some things it wants to have or do, and others it wants to avoid.' 'It prefers some things to others.' 'It does this with the intention of doing that.' 'At first it did not

intend to do this, but after deliberating about the alternatives it changed its mind.' 'There are some things it is aware of, others it is not aware of (e.g. in its environment) at any time.' Further, explanations of its actions could be given in terms of beliefs and desires, for instance, 'It pushed the chair from the doorway because it wanted to replace its batteries and believed that a stock of fresh batteries was in a box in the next room, and that it could only get to the next room by moving the chair then going through the door.' However, the physical events and processes involved in all this might be too complex for an explanation of the physical behaviour in terms of current physical theory to be a practical possibility, even though it was known from the initial design and construction that the system conformed to thesis (P^1), i.e. it was a physical system.

New levels of sophistication in the types of action it can perform, and in the kinds of mental phenomena which can be ascribed to it would arise out of the satisfaction by the system of a further condition:

9. It can use a language to communicate with us or with others like itself about the contents of the environment, its own internal states and processes, its goals, its unsolved problems, etc. That is, it can translate from an internal representation into an external language and vice versa, and can use linguistic devices to indicate whether it is requesting information, reporting something it accepts as correct, formulating a hypothesis for discussion, etc., i.e. illocutionary-act indicators.

If this additional condition were satisfied, we could talk about telling it things, persuading or advising it, bargaining with it (e.g. 'If you will carry this box across the road for me, I'll re-solder your faulty connections') etc. We could influence its behaviour in these ways, that is by influencing its beliefs or desires, even though the underlying physical processes were too complex for us to comprehend. It would clearly no longer be true to say of such a system that all it does is what its designer intended it to do. Several such systems interacting with one another could form a community engaged in various co-operative enterprises and develop various conventions for dealing with possible or actual

conflicts (e.g. law courts) or for making conflicts less likely (e.g. traffic regulations), etc. Confronted with an autonomous community of such robots, we should surely find it intolerable to sustain the clumsy circumlocutions some people would at first want to use in describing and explaining their behaviour because they know 'those things are only physical mechanisms'. There seems to me to be no conceptual mistake involved in thinking that it would be *morally* wrong not to regard such things as conscious, intelligent, responsible for their actions and worthy of being treated with consideration for their desires and interests.

Of course, I cannot yet rigorously *prove* that a physical system could satisfy the conditions (1)—(9), or even conditions (1)—(8): that would require me to design a system satisfying these requirements, or else to prove that the human (or some animal) brain is a physical system. I doubt if anyone can yet design such a system, though the development of computer science holds out encouraging prospects. (The formulation of some of the conditions was influenced by a paper by Max Clowes, 'On seeing things' in *Artificial Intelligence* Vol. 2, 1971, describing a computer programme for the interpretation of pictures of polyhedra.)

Anyone who wished to argue against the position in this paper must argue that a physical system could not satisfy the conditions listed, or that the conditions are not *sufficient* to justify the ascription of psychological predicates and explanations. To establish the latter would involve specifying some *necessary* condition which is not entailed by the conjunction of my conditions.

Unfortunately, the issue is in part bedevilled by the indeterminacy of our concepts. For instance, if anyone claims that an additional condition is being a member of some biological species which has evolved naturally, which would rule out artefacts, then I cannot argue against this except to say (a) that my own concepts have no such limitations and (b) that situations could arise in which insisting on this condition as necessary would lead to easily avoidable terminological complexity. I could also try to show that the restriction would be immoral, but a contrary moral position, that artefacts ought not to be treated as conscious

being with beliefs, etc., can be consistently maintained, as can a similar moral attitude to dogs, cats, slaves, or members of the 'lower' classes. However, apart from the moral implications, a disagreement on whether the biological condition is necessary for the applicability of psychological concepts seems to have no philosophical interest: it seems to be a purely terminological disagreement. Further, the intelligibility, to many readers, of science fiction stories in which such concepts are applied to robots of various sorts seems to me to show that I am in large company in *not* finding the biological condition logically or conceptually necessary.

7. SHOULD THE LIST OF CONDITIONS MENTION 'INNER EXPERIENCE'?

My list of conditions does not explicitly include the possession of 'inner experience', the sort of content of consciousness that we are 'directly aware of, etc. This, it could be argued, shows that the conditions are not sufficient to justify talk of mental phenomena. Part of the reply to this objection is that the claim that robots satisfying all my conditions also have this inner world would be no more (and no less) problematic than the claim that other human beings do. No doubt some such robots of the future will be convinced that they are not mere physical systems since they do have this 'something extra' which they can identify for themselves by focusing attention inwards. There is no conceptual difficulty in supposing that such a robot might learn that its sensory systems can mislead it, and that it can be given hallucinations by a malevolent human being: and this might lead it along the well-worn track to the conclusion that the only thing it can be sure of is the content of its own consciousness, since even the supposed existence of its body and an external world could be merely a complex illusion.

But there are further subtleties underlying the appeal to this extra condition. We are able intelligibly to talk about our private inner experience, the contents of our sensations, etc., as inaccessible to other people because of the following facts. Whatever I can refer to it is possible for someone else to refer to also. However, the manner in which I am able to identify it (its 'mode of presentation' to me, in Frege's terminology)

need not be the same as someone else's. I have my own point of view, and I therefore experience only certain aspects of the thing and its relations to the rest of the world. (This is what makes it possible for Frege to distinguish the *Sinn* of a referring expression from its *Bedeutung*. See 'The Thought, a Logical Enquiry', in *Philosophical Logic*, ed. P. F. Strawson, for Frege's most mature published thought on this topic.) However, I can, and you can, refer to my point of view, the aspects I am aware of, etc. That is, we can both refer to the object's 'mode of presentation' to me. But this again will be something I refer to from a different viewpoint: if Z is the original thing referred to, and Y is its mode of presentation to me, then Y, like Z can be referred to by other people, but, like Z, Y will also have a unique mode of presentation to me, call it X. Again, I can refer to X, and so can you. By always insisting that whatever we refer to or discuss publicly, there is always something more, a point of view, an aspect, a mode of presentation, uniquely underlying *my* manner of referring to it, I adopt what might be called the strategy of *always pointing nearer to self*. The idea of a private, inner, domain, accessible to nobody but oneself and identifiable to oneself by a 'private ostensive definition', seems to be an idea arrived at by postulating a limiting case defined by successive application of this strategy. Note that I am not saying that this concept is confused, incoherent, or whatever: I regard it as an important fact about our language, thought and experience that it makes this strategy possible, and any philosophical system or theory of language which rules it out is to that extent descriptively inadequate. (Even the limiting case can perhaps be made respectable by defining it as the union of all the 'modes of presentation' of objects to the person in question.)

However, this very strategy is available to a robot satisfying my conditions and able to represent, and therefore think about, its relation to the things it perceives and refers to. So the possession of a private world of the sort under discussion is, after all, entailed by the satisfaction of conditions like the ones I have formulated, even though it is not mentioned explicitly.

At this point some philosophers will argue that my

discussion fails to give an adequate account of the *ontological status* of a mind and its contents. I can only reply that my attempts to follow discussions of ontological status have so far left me completely unable to grasp what the problem is, that is, unable to see what there is to talk about, over and above the sorts of issues I have discussed here.

8. CONCLUSION

I have tried to show (1) that the thesis (P) is empirical and therefore not amenable to philosophical refutation; (2) that it does not imply that mental phenomena are identical with, or composed of, or reliably inductively correlated with physical phenomena; (3) that it does not imply that human actions have purely physical explanations; (4) that it does not imply that mental events and processes can have no causal influence on our actions. In addition, I have tried to show how it is possible for our psychological concepts to get a grip on a physical system, by describing conditions which justify their application and which could be satisfied by a physical system. In particular, I have tried to show how such a system might, like a human being, have its own private, inner experience.

I originally hoped to conclude with a discussion of the claim that (P) implies that all deliberation is pointless or impossible, that there is no such thing as moral responsibility, and that moral assessment of human actions is impossible. I do not think that (P) implies any of these things, but this paper is already too long, so I leave these topics untouched. No doubt readers will be able to predict how I would deal with them.

Note added 29 Dec 2005:

The original publication included a response to my paper, written before the conference, by two psychologists, George Mandler (UCSD) and William Kessen (Yale), entitled 'The appearance of free will' (pages 305 to 324), followed by a commentary by Alan R. White (Philosopher, Hull University), entitled 'Chairman's Remarks' (pages 325 to 330). Then followed (pages 331 to 339) several short discussion points submitted by people attending the conference (Sloman, Philippa Foot, Anita Gregory, Les Holborrow, Donald MacKay, Robin Attfield, Sloman), and finally a set of Concluding Remarks by Mandler and Kessen (pages 340 to 342) and Sloman (pages 343-7).

My own contributions to the discussion are below. I have not included the contributions of other authors partly because of copyright worries and partly for lack of time. However it is worth noting that Alan White defended the mind-brain identity thesis against my argument that the relation between intentional action and the corresponding physical process was a three-termed relation, and I accept this criticism in my discussion, though not the identity thesis.

Discussion

DR SLOMAN

Professor White is perfectly correct in finding fault with my attempt to prove that actions and mental processes were neither identical with nor composed of physical behaviour. In essence, my argument went: 'there is a three-termed relation between actions and physical behaviour, whereas identity and composition are two-termed relations, so actions and physical behaviour are not identical, etc.' This argument is not only fallacious, but stupid, since it assumes that two objects cannot be involved simultaneously in a two-termed and in a three-termed relation. I can only apologise for wasting the time of readers of the original paper over this point.

The source of my error was the tendency to talk of '*The relation*' as if only one relation existed between actions and physical behaviour. Clearly there are many different relations, as White observes in his paragraph (a). Yet he continues to exhibit the same tendency as led me into error, for instance in the opening sentences of his paragraphs 1) and 2).

Note added 29 Dec 2005:

White insists in his paragraph 1) that there is a distinction between behaviour of *people* including blushing, gasping, and believing, and behaviour of *bodies*. In his paragraph 2) he makes the important point that there is not just *one* mind-body relation but a variety of different relationships that need to be studied piecemeal. E.g. he suggests that the relation between sensations and brain processes may be unlike the relation between thoughts and neurological processes. This is very close to the point I was trying to make.

Various commentaries followed my note which are omitted here.

DR SLOMAN:

Mr Attfield's comments are based on the assumption that I was talking about freedom, (which I don't think I ever mentioned), and that I was trying to decide whether sometimes human agents could have acted otherwise, which I regard as an ill-formed problem. In order to refute my supposed claims on these matters, he tries to argue that if physicalism (as I defined it) is true, then 'the psychological antecedents of my body's movements are thus redundant in the prediction and explanation of those movements'. He does not appear to realise that this conclusion is little more than a reformulation of the physicalist premise presented as an empirical hypothesis in section B of my paper. The premise states that all physical behaviour of human bodies (and their parts) conforms to current physical theory. I wrote that this was meant to imply that in so far as the physically describable movements can be explained and predicted they can be explained and predicted in purely physical terms. So psychological antecedents are obviously redundant in such explanations and predictions. This is far from being incompatible with anything I asserted. It does, however, leave open the possibility, mentioned in the fourth last paragraph of my section D, that there might be a different set of premises, equally capable of predicting and explaining those movements, in which the mention of psychological antecedents was not redundant.

It seems that Attfield's main problem is whether human agents sometimes could have acted otherwise. The answer is

obviously 'yes, they could have', just as it is obvious that some physical mechanism could have behaved otherwise. For instance, there could have been more petrol in the tank, in which case the engine would not have spluttered to a halt when it did. To produce a serious philosophical problem about whether human agents could have acted otherwise, one must specify a set of conditions supposed to remain fixed. But what conditions? If physicalism (as defined in my paper) is correct, then in theory one can always find some physical specification of a set of conditions from which one can infer with the aid of physical theories that (apart from quantum indeterminacy) no physical process could have occurred in those conditions other than what did occur. It is not clear whether Mr Atfield finds this conclusion unpalatable or rejoices in it, but it certainly seems to disturb some people and I have tried to diagnose the unpalatability as due, in part, to a mistaken assumption that the conclusion implies that human actions can be fully explained in physical terms and that beliefs, hopes, wants, fears, moral ideals and the like cannot influence actions. My paper attempted to refute the assumption. (The argument of my paper could be rejected by showing that Professor White's enormously hospitable empirical identity relation transmits explanatory force without any explicit additional premises being needed. This far from obvious.)

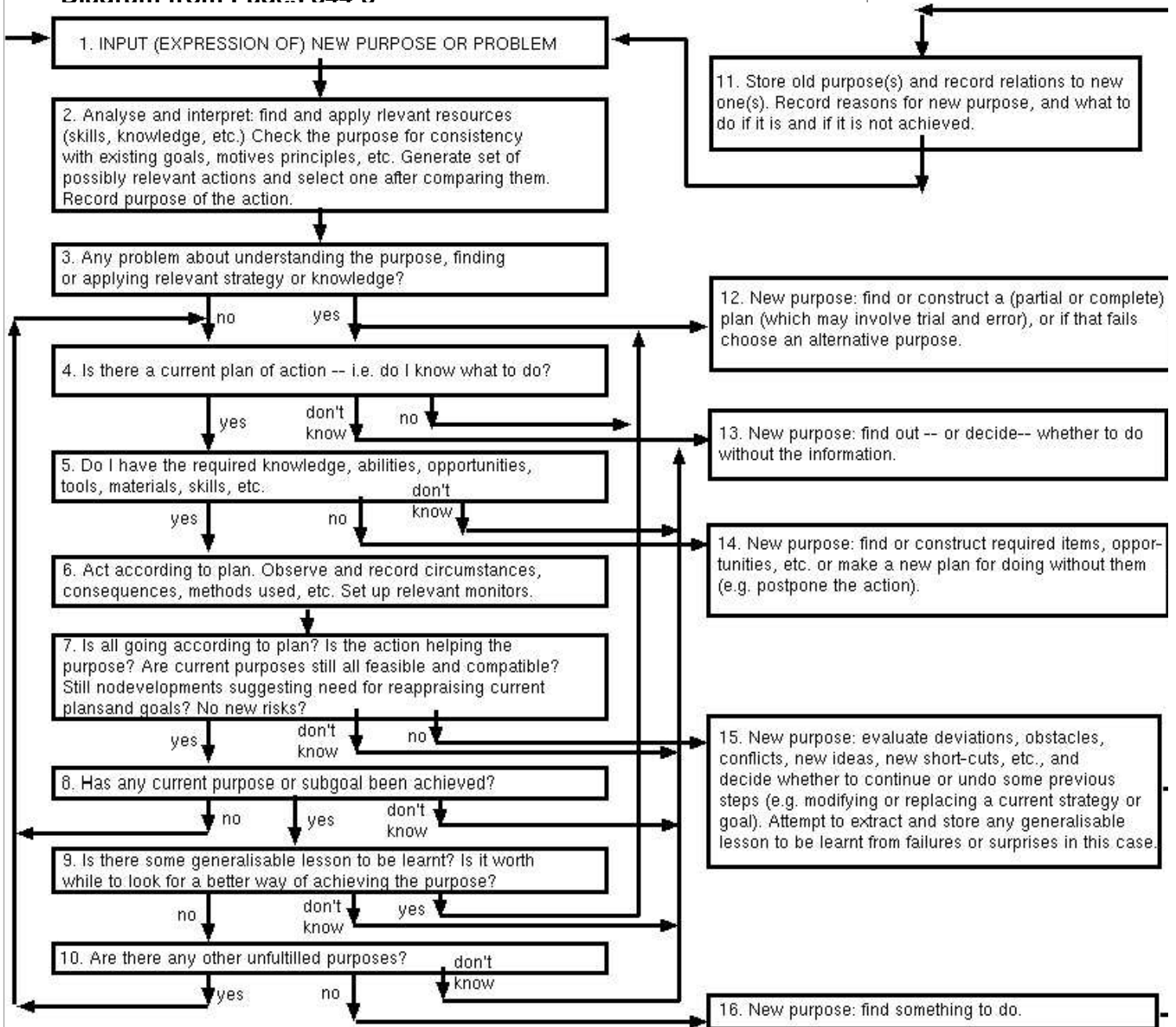
DR SLOMAN:

Finally, I should like to point out that I grow increasingly convinced that all debates about determinism, free will and the relations between mind and body, between actions and movements, between mental processes and brain processes, and so on are a waste of time until we have a much clearer and more detailed and systematic characterisation of what sorts of things agents, actions, decisions, thoughts, beliefs, desires, and so on are. This characterisation will not come from introspection, nor from piecemeal hacking away at the puzzles that analytical philosophers are so fond of, nor from psychological experiments, but from a sustained effort to arrive at a new more perspicuous representation of a human being which accounts for the *fine-structure* and the *variety* of the enormous range of facts about human abilities and possibilities that we all know and constantly take for granted in our daily social interactions. These common-sense facts have to be collected and organised in a systematic way so as to provide a basis for constructing and testing a theory of the kind of mechanism a human mind is.

The most penetrating advances in the development of powerful new means of representing aspects of mind are being made by those who attempt to formulate their theories in computer programmes. Practitioners of the discipline of *Artificial Intelligence* have been forced to develop and test all sorts of new tools for representing aspects of mental abilities, and the study of its literature is indispensable for anyone interested in characterising human mental abilities. (One of the most important, and most accessible, contributions so far is T. Winograd's M.I.T. Ph.D. thesis, 'Procedures as a representation for data in a computer program for understanding natural language'. This has appeared in the journal *Cognitive Psychology*, January 1972, and has been published as a book by Academic Press and Edinburgh University Press.)

Of course, such programming models will be inadequate in all sorts of ways for some time yet, but the inadequacies are rapidly discovered and (not quite so rapidly) attended to. When we have a better idea of what sorts of things can be

Diagram from Pages 344-5



achieved by computer programmes (a new type of mechanism) and how they are related to the computers which contain them (an old type mechanism), we shall be in a much better position to ask sensible questions about the relations between mental and bodily phenomena, and about free will and determinism.

For instance, a not very deep analysis of common-sense knowledge about what sorts of processes a person can go through in performing a variety of tasks, including solving problems, playing games, constructing models from a 'Meccano' kit, explaining accidents, or reading a letter in obscure handwriting; together with analysis of the meanings of words and phrases like 'careful', 'rash', 'reckless', 'attentive', 'alert', 'done from habit', 'learning from experience', 'trial and error', 'bear in mind', 'intended', 'intentional', etc., reveals that even in fairly mundane forms of human activity the range of possibilities for mental processes is at least as rich as the range represented by all the routes through the accompanying flow-chart (Fig. 1). Moreover, which of the possible routes is taken at any stage is normally determined by the sorts of processes crudely indicated by the wording in the boxes, yet also subject to higher levels of control based on an awareness of what is going on and the results of tests carried out by the 'monitors' mentioned in box 2. (Psychologists please note: this chart is not offered as a new psychological theory, merely as a convenient and fairly economical summary of a large number of fairly obvious common-sense facts.)

It can be argued that a mechanism generating at least the range of possibilities expressed in the flow chart (with all sorts of further complexities hidden in the boxes of the chart) is minimally required for any system which can behave with the kind of intelligence quite obviously exhibited by human beings of all ages and many other animals (e.g. the chimpanzees described in W. Kohler, *The Mentality of Apes*). For a system with this range of possibilities inherent in it there is obviously ample application for the concept 'could have done otherwise'. If this is what freedom of the will is about, then any intelligent system must have freedom of the will. In this sense talk of freedom of the will is by no means

'inherited word play' that 'can be safely put aside' as Mandler and Kessen suggest. Older exemplars of mechanisms, such as the solar system, steam engines, and control mechanisms with feedback loops, were clearly unable to generate this rich kind of range of possibilities and so it used to seem obvious that nothing describable as a mechanism could underly the kind of thing we all know the human mind to be; and moreover any suggestion that such mechanisms might adequately model our minds was unpalatable on account of the implied restrictions on possibilities for human choice. The only previously known mechanisms capable of generating a rich enough range of possibilities were meaningless agglomerations of randomly related items, like the molecules of a gas or perhaps a vast collection of roulette wheels. There was always therefore an unhappy tension between excessively rigid deterministic models and excessively purposeless random models.

No computing system known to me has the rich range of possibilities for purposive action expressed in the chart, though a superficial reading of Winograd's thesis mentioned above may give the impression that his programme has. However, it seems clear that programming languages, with their facilities for expressing conditions under which different processes occur, and computing systems with their potential for rapid, goal-directed, changes of complex internal structures, together provide new tools for thinking about, and building and testing, new mechanisms which avoid both extremes. Among the problems still to be solved, however, is the problem of finding means of expressing and storing vast ranges of very varied kinds of 'knowledge' in forms which make them readily accessible when they are relevant to current purposes and contexts and also readily modifiable in the light of new information or the discovery of internal inconsistencies, etc. (Compare boxes 2, 8 and 15 of the flow-chart.) Here lie rich new pastures for philosophers and psychologists interested in concepts like 'belief', 'skill', 'habit', 'association of ideas', 'learning', 'memory', etc.

To argue in advance that such attempts to represent human mental abilities are futile because of the nature of the physical processes known to underly computing systems, as

Herbert Dreyfus attempted to do in his paper, is, in the current state of ignorance about what can and cannot be programmed, like arguing that human brains could not possibly provide a basis for human behaviour because of the nature of the atoms and molecules of which they are composed.

Similarly, in the current state of ignorance about what sorts of powers can or cannot be programmed into computers, about 99 per cent of philosophical discussion of problems about the relation between mind and body and about the extent to which human actions are or are not determined is pointless.¹ This applies to my own paper.

¹ Several of the points made here can be found in Marvin Minsky's introduction to *Semantic Information Processing*, edited by him. (M.I.T. Press, 1968)