18

Minds and machines*

The various issues and puzzles that make up the traditional mind–body problem are wholly linguistic and logical in character: whatever few empirical 'facts' there may be in this area support one view as much as another. I do not hope to establish this contention in this paper, but I hope to do something toward rendering it more plausible. Specifically, I shall try to show that all of the issues arise in connection with any computing system capable of answering questions about its own structure, and have thus nothing to do with the unique nature (if it *is* unique) of human subjective experience.

To illustrate the sort of thing that is meant: one kind of puzzle that is sometimes discussed in connection with the 'mind–body problem' is the puzzle of *privacy*. The question 'How do I know I have a pain?' is a *deviant*† ('logically odd') question. The question 'How do I know Smith has a pain?' is not all at deviant. The difference can also be mirrored in impersonal questions: 'How does anyone ever know he himself has a pain?' is deviant; 'How does anyone ever know that someone else is in pain?' is non-deviant. I shall show that the difference in status between the last two questions is mirrored in the case of machines: if T is a *Turing machine* (see below), the question 'How does T ascertain that it is in state A?' is, as we shall see, 'logically odd' with a vengeance; but if T is capable of investigating its neighbor machine T' (say, T has electronic 'sense-organs' which 'scan' T'), the question 'How does T ascertain that T' is in state A?' is not at all odd.

Another question connected with the 'mind–body problem' is the question whether or not it is ever permissible to identify mental events and physical events. Of course, I do not claim that this question arises for Turing machines, but I do claim that it is possible to construct a logical analogue for this question that does arise, and that all of the question of 'mind–body identity' can be mirrored in terms of the analogue.

* First published in Sidney Hook (ed.) *Dimensions of Mind* (New York, 1960). Reprinted by permission of New York University Press.

† By a 'deviant' utterance is here meant one that deviates from a semantical regularity (in the appropriate natural language). The term is taken from Ziff, 1960.

To obtain such an analogue, let us identify a scientific theory with a 'partially-interpreted calculus' in the sense of Carnap†. Then we can perfectly well imagine a Turing machine which generates theories, tests them (assuming that it is possible to 'mechanize' inductive logic to some degree), and 'accepts' theories which satisfy certain criteria (e.g. predictive success). In particular, if the machine has electronic 'sense organs' which enable it to 'scan' itself while it is in operation, it may formulate theories concerning its own structure and subject them to test. Suppose the machine is in a given state (say, 'state A') when, and only when, flip-flop 36 is on. Then this statement: 'I am in state A when, and only when, flip-flop 36 is on', may be one of the theoretical principles concerning its own structure accepted by the machine. Here 'I am in state A' is, of course, 'observation language' for the machine, while 'flip-flop 36 is on' is a 'theoretical expression' which is partially interpreted in terms of 'observables' (if the machine's 'sense organs report by printing symbols on the machine's input tape, the 'observables' in terms of which the machine would give a partial operational definition of 'flip-flop 36 being on' would be of the form 'symbol # so-and-so appearing on the input tape'). Now all of the usual considerations for and against mind–body identification can be paralleled by considerations for and against saying that state A is in fact *identical* with flip-flop 36 being on.

Corresponding to Occamist arguments for 'identify' in the one case are Occamist arguments for identity in the other. And the usual argument for dualism in the mind–body case can be paralleled in the other as follows: for the machine, 'state A' is directly observable; on the other hand, 'flip-flops' are something it knows about only via highly-sophisticated inferences – How *could* two things so different *possibly* be the same? This last argument can be put into a form which makes it appear somewhat stronger. The proposition:

(1) I am in state A if, and only if, flip-flop 36 is on,

is clearly a 'synthetic' proposition for the machine. For instance, the machine might be in state A and its sense organs might report that flip-flop 36 was *not* on. In such a case the machine would have to make a methodological 'choice' – namely, to give up (1) or to conclude that it had made an 'observational error' (just as a human scientist would be confronted with similar methodological choices in studying his own

†Cf. Carnap 1953 and 1956. This model of a scientific theory is too oversimplified to be of much general utility, in my opinion: however, the oversimplifications do not affect the present argument.

psychophysical correlations). And just as philosophers have argued from the synthetic nature of the proposition:

(2) I am in pain if, and only if, my C-fibers are stimulated,

to the conclusion that the *properties* (or 'states' or 'events') being in pain, and having C-fibers stimulated, cannot possibly be the same (otherwise (2) would be analytic, or so the argument runs); so one should be able to conclude from the fact that (1) is synthetic that the two properties (or 'states' or 'events') – being in state A and having flip-flop 36 on – cannot possibly be the same!

It is instructive to note that the traditional argument for dualism is not at all a conclusion from 'the raw data of direct experience' (as is shown by the fact that it applies just as well to non-sentient machines), but a highly complicated bit of reasoning which depends on (a) the reification of universals† (e.g. 'properties', 'states', 'events'); and on (b) a sharp analytic-synthetic distinction.

I may be accused of advocating a 'mechanistic' world-view in pressing the present analogy. If this means that I am supposed to hold that machines think,‡ on the one hand, or that human beings are machines, on the other, the charge is false. If there is some version of mechanism sophisticated enough to avoid these errors, very likely the considerations in this paper support it.§

1. Turing Machines

The present paper will require the notion of a *Turing machine*‖ which will now be explained.

Briefly, a Turing machine is a device with a finite number of internal configurations, each of which involves the machine's being in one of a finite number of *states*,¶ and the machine's scanning a tape on which certain symbols appear.

† This point was made by Quine in Quine, 1957.

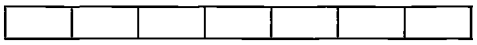
‡ Cf. Ziff's paper (1959) and the reply (1959) by Smart. Ziff has informed me that by a 'robot' he did not have in mind a 'learning machine' of the kind envisaged by Smart, and he would agree that the considerations brought forward in his paper would not necessarily apply to such a machine (if it can properly be classed as a 'machine' at all). On the question of whether 'this machine thinks (feels, etc.)' is *deviant* or not, it is necessary to keep in mind both the point raised by Ziff (that the important question is not whether or not the utterance is deviant, but whether or not it is deviant for non-trivial reasons), and also the 'diachronic–synchronic' distinction discussed in section 5 of the present paper.

§ In particular, I am sympathetic with the general standpoint taken by Smart in (1959b) and (1959c). However, see the linguistic considerations in section 5.

‖ For further details, cf. Davis, 1958 and Kleene, 1952.

¶ This terminology is taken from Kleene, 1952, and differs from that of Davis and Turing.

The machine's tape is divided into separate squares, thus:



on each of which a symbol (from a fixed finite alphabet) may be printed. Also the machine has a 'scanner' which 'scans' one square of the tape at a time. Finally, the machine has a *printing mechanism* which may (a) *erase* the symbol which appears on the square being scanned, and (b) print some other symbol (from the machine's alphabet) on that square.

Any Turing machine is completely described by a *machine table*, which is constructed as follows: the rows of the table correspond to letters of the alphabet (including the 'null' letter, i.e. blank space), while the columns correspond to states *A*, *B*, *C*, etc. In each square there appears an 'instruction', e.g. '$s_5L\ A$', '$s_7C\ B$', '$s_3R\ C$'. These instructions are read as follows: '$s_5L\ A$' means 'print the symbol s_5 on the square you are now scanning (after erasing whatever symbol it now contains), and proceed to scan the square immediately to the left of the one you have just been scanning; also, shift into state *A*.' The other instructions are similarly interpreted ('*R*' means 'scan the square immediately to the *right*', while '*C*' means 'center', i.e. continue scanning the *same* square). The following is a sample machine table:

		A	*B*	*C*	*D*
(s_1)	1	s_1RA	s_1LB	s_3LD	s_1CD
(s_2)	+	s_1LB	s_2CD	s_2LD	s_2CD
(s_3)	blank space	s_3CD	s_3RC	s_3LD	s_3CD

The machine described by this table is intended to function as follows: the machine is started in state *A*. On the tape there appears a 'sum' (in unary notion) to be 'worked out', e.g. '11 + 111.'

The machine is initially scanning the first '1'. The machine proceeds to 'work out' the sum (essentially by replacing the plus sign by a 1, and then going back and erasing the first 1). Thus if the 'input' was 1111 + 11111 the machine would 'print out'111111111, and then go into the 'rest state' (state *D*).

A 'machine table' *describes* a machine if the machine has internal states corresponding to the columns of the table, and if it 'obeys' the instruction in the table in the following sense: when it is scanning a square on which a symbol s_1 appears and it is in, say, state *B*, that it carries out the 'instruction' in the appropriate row and column of the table (in this case, column *B* and row s_1). Any machine that is described by a machine table of the sort just exemplified is a Turing machine.

The notion of a Turing machine is also subject to generalization† in various ways – for example, one may suppose that the machine has a second tape (an 'input tape') on which additional information may be printed by an operator in the course of a computation. In the sequel we shall make use of this generalization (with electronic 'sense organs' taking the place of the 'operator').

It should be remarked that Turing machines are able in principle to do anything that any computing machine (of whichever kind) can do.‡

It has sometimes been contended (e.g. by Nagel and Newman in their book *Gödel's Proof*) that 'the theorem [i.e. Gödel's theorem] does indicate that the structure and power of the human mind are far more complex and subtle than any non-living machine yet envisaged' (p. 10), and hence that a Turing machine cannot serve as a model for the human mind, but this is simply a mistake.

Let T be a Turing machine which 'represents' me in the sense that T can prove just the mathematical statements I can prove. Then the argument (Nagel and Newman give no argument, but I assume they must have this one in mind) is that by using Gödel's technique I can discover a proposition that T cannot prove, and moreover I can prove this proposition. This refutes the assumption that T 'represents' me, hence I am not a Turing machine. The fallacy is a misapplication of Gödel's theorem, pure and simple. Given an arbitrary machine T, all I can do is find a proposition U such that I can prove:

(3) If T is consistent, U is true,

where U is undecidable by T if T is in fact consistent. However, T can perfectly well prove (3) too! And the statement U, which T *cannot* prove (assuming consistency), I cannot prove either (unless I can prove that T is consistent, which is unlikely if T is very complicated)!

2. Privacy

Let us suppose that a Turing machine T is constructed to do the following. A number, say '3000' is printed on T's tape and T is started in T's 'initial state'. Thereupon T computes the 3000th (or whatever the given number was) digit in the decimal expansion of π, prints this digit on its tape, and goes into the 'rest state' (i.e. turns itself off).

† This generalization is made in Davis, 1958, where it is employed in defining relative recursiveness.

‡ This statement is a form of *Church's thesis* (that recursiveness equals effective computability).

Clearly the question 'How does T "ascertain" [or "compute", or "work out"] the 3000th digit in the decimal expansion of π?' is a sensible question. And the answer might well be a complicated one. In fact, an answer would probably involve three distinguishable constituents:

(i) A description of the sequence of states through which T passed in arriving at the answer, and of the appearance of the tape at each stage in the computation.

(ii) A description of the *rules* under which T operated (these are given by the 'machine table' for T).

(iii) An explanation of the *rationale* of the entire procedure.

Now let us suppose that someone voices the following objection: 'In order to perform the computation just described, T must pass through states A, B, C, etc. But how can T ascertain that it is in states A, B, C, etc?'

It is clear that this is a silly objection. But what makes it silly? For one thing, the 'logical description' (machine table) of the machine describes the state only in terms of their *relations* to each other and to what appears on the tape. The 'physical realization' of the machine is immaterial, so long as there *are* distinct states A, B, C, etc., and they succeed each other as specified in the machine table. Thus one can answer a question such as 'How does T ascertain that X?' (or 'compute X', etc.) only in the sense of describing the *sequence of states* through which T must pass in ascertaining that X (computing, X etc.), the rules obeyed, etc. But there is no 'sequence of states' through which T must pass to be in a single state!

Indeed, suppose there were – suppose T could not *be* in state A without first *ascertaining* that it was in state A (by first passing through a sequence of other states). Clearly a vicious regress would be involved. And one 'breaks' the regress simply by noting that the machine, in ascertaining the 3000th digit in π, *passes through* its states – but it need not in any significant sense 'ascertain' that it is passing through them.

Note the analogy to a fallacy in traditional epistemology: the fallacy of supposing that to know that p (where p is any proposition) one must first know that q_1, q_2, etc. (where q_1, q_2, etc., are appropriate *other* propositions). This leads either to an 'infinite regress' or to the dubious move of inventing a special class of 'protocol' propositions.

The resolution of the fallacy is also analogous to the machine case. Suppose that on the basis of sense experiences E_1, E_2, etc., I know that there is a chair in the room. It does not follow that I verbalized (or even *could* have verbalized) E_1, E_2, etc., nor that I remember E_1, E_2, etc., nor

even that I 'mentally classified' ('attended to', etc.) sense experiences E_1, E_2, etc., when I had them. In short, it is necessary to *have* sense experiences, but not to *know* (or even *notice*) what sense experiences one is having, in order to have certain kinds of knowledge.

Let us modify our case, however, by supposing that whenever the machine is in one particular state (say, 'state A') it prints the words 'I am in state A'. Then someone might grant that the machine does not in general ascertain what state it is in, but might say in the case of state A (after the machine printed 'I am in state A'): 'The machine ascertained that it was in state A'.

Let us study this case a little more closely. First of all, we want to suppose that when it is in state A the machine prints 'I am in state A' without first passing through any other states. That is, in every row of the column of the table headed 'state A' there appears the instruction: *print*† '*I am in state* A'. Secondly, by way of comparison, let us consider a human being, Jones, who says 'I am in pain' (or 'Ouch!', or 'Something hurts') whenever he is in pain. To make the comparison as close as possible, we will have to suppose that Jones' linguistic conditioning is such that he simply says 'I am in pain' 'without thinking', i.e. without passing through any introspectible mental states other than the pain itself. In Wittgenstein's terminology, Jones simply *evinces* his pain by saying 'I am in pain' – he does not first reflect on it (or heed it, or note it, etc.) and then consciously describe it. (Note that this simple possibility of uttering the 'proposition', 'I am in pain' without first performing any mental 'act of judgement' was overlooked by traditional epistemologists from Hume to Russell!) Now we may consider the parallel question 'Does the machine "ascertain" that it is in state A?' and 'Does Jones "know" that he is in pain?' and their consequences.

Philosophers interested in semantical questions have, as one might expect, paid a good deal of attention to the verb 'know'. Traditionally, three elements have been distinguished: (1) 'X know that p' implies that p is *true* (we may call this the *truth* element); (2) 'X knows that p' implies that X believes that p (philosophers have quarrelled about the word, some contending that it should be 'X is *confident* that p,' or 'X *is in a position to assert* that p'; I shall call this element the *confidence* element); (3) 'X knows that p' implies that X has evidence that p (here I think the word 'evidence' is definitely wrong,‡ but it will not matter for present purposes; I shall call this the *evidential* element). Moreover,

† Here it is necessary to suppose that the entire sentence 'I am in state A.' counts as a single symbol in the machine's alphabet.

‡ For example, I know that the sun is 93 million miles from the earth, but I have no *evidence* that this is so. In fact, I do not even remember where I learned this.

it is part of the meaning of the word 'evidence' that nothing can be literally evidence for itself: if X is evidence for Y, then X and Y must be different things.

In view of such analyses, disputes have arisen over the propriety of saying (in cases like the one we are considering) 'Jones knows that he is in pain.' On the one hand, philosophers who take the common-sense view ('When I have a pain I *know* I have a pain') argue somewhat as follows: it would be clearly false to say Jones does *not* know he has a pain; but either Jones knows or he does not; hence, Jones knows he has a pain. Against these philosophers, one might argue as follows: 'Jones does not know X' implies Jones is not in a position to assert that X; hence, it is certainly wrong to say 'Jones does not know he has a pain.' But the above use of the Law of the Excluded Middle was fallacious: words in English have *significance ranges*, and what is contended is that it is not semantically correct to say *either* 'Jones knows that he has a pain.' *or* 'Jones does not know he has a pain.' (although the former sentence is certainly less misleading than the latter, since *one* at least of the conditions involved in knowing is met – Jones is in a position to assert he has a pain. (In fact the *truth* and *confidence* elements are both present; it is the evidential element that occasions the difficulty.)

I do not wish to argue this question here;† the present concern is rather with the similarities between our two questions. For example, one might decide to accept (as 'non-deviant', 'logically in order', 'non-selfcontradictory', etc.) the two statements:

(a) The machine ascertained that it was in state A,
(b) Jones knew that he had a pain,

or one might reject both. If one rejects (a) and (b), then one can find alternative formulations which are certainly semantically acceptable: e.g. (for (a)) 'The machine was in state A, and this caused it to print: "I am in state A"', (for (b)) 'Jones was in pain, and this caused him to say "I am in pain"' (or, 'Jones was in pain, and he evinced this by saying "I am in pain"').

On the other hand, if one accepts (a) and (b), then one must face the questions (a_1) '*How* did the machine ascertain that it was in state A?', and (b_1) '*How* did Jones know that he had a pain?'

And if one regards these questions as having answers at all, then they

† In fact, it would be impossible to decide whether 'Jones knows he has a pain' is deviant or not without first reformulating the evidential condition so as to avoid the objection in note ‡ on p. 368 (if it can be reformulated so as to save anything of the condition at all). However the discussion above will indicate, I believe, why one might *want* to find that this sentence is deviant.

will be degenerate answers – e.g. 'By being in state A' and 'By having the pain.'

At this point it is, I believe, very clear that the difficulty has in both cases the same cause. Namely, the difficulty is occasioned by the fact that the 'verbal report' ('I am in state A', or 'I am in pain') issues directly from the state it 'reports': no 'computation' or additional 'evidence' is needed to arrive at the 'answer'. And the philosophic disagreements over 'how to talk' are at bottom concerned with finding a terminology for describing cognitive processes in general that is not misleading in this particular case. (Note that the traditional epistemological answer to (b_1) – namely, 'by introspection' – is false to the facts of this case, since it clearly implies the occurrence of a mental event (the 'act' of introspection) distinct from the feeling of pain.)

Finally, let us suppose that the machine is equipped to 'scan' its neighbor machine T_1. Then we can see that the question 'How does T ascertain that T_1 is in state A?' may be a perfectly sensible question, as much so as 'How does T ascertain that the 3000th digit of π is so-and-so?' In both cases the answer will involve describing a whole 'program' (plus explaining the *rationale* of the program, if necessary). Moreover, it will be necessary to say something about the physical context linking T and T_1 (arrangement of sense organs, etc.), and not just to describe the internal states of T: this is so because T is now answering an *empirical* and not a mathematical question. In the same way 'How did Sherlock Holmes know that Jones was in pain?' may be a perfectly sensible question, and may have quite a complicated answer.

3. 'Mental' states and 'logical' states

Consider the two questions:

(1) How does Jones know he has a pain?
(2) How does Jones know he has a fever?

The first question is, as we saw in the preceding section, a somewhat peculiar one. The second question may be quite sensible. In fact, if Jones says 'I have a pain' no one will retort 'You are mistaken'. (One *might* retort 'You have made a slip of the tongue' or 'You are lying', but not 'You are *mistaken*'.) On the other hand, if Jones says 'I have a fever', the doctor who has just taken Jones' temperature may quite conceivably retort 'You are mistaken'. And the doctor need not mean that Jones made a linguistic error, or was lying, or confused.

It might be thought that, whereas the difference between statements about one's own state and statements about the states of others has an

analogue in the case of machines, the difference, just touched upon, between statements about one's 'mental' state and statements about one's 'physical' state, in traditional parlance, does not have any analogue. But this is not so. Just what the analogue is will now be developed.

First of all, we have to go back to the notion of a Turing machine. When a Turing machine is described by means of a 'machine table', it is described as something having a tape, a printing device, a 'scanning' device (this may be no more than a point of the machine which at any given time is aligned with just one square of the tape), and a finite set (*A*, *B*, *C*, etc.) of 'states'. (In what follows, these will be referred to at times as *logical states* to distinguish them from certain other states to be introduced shortly.) Beyond this it is described only by giving the deterministic rules which determine the order in which the states succeed each other and what is printed when.

In particular, the 'logical description' of a Turing machine does not include any specification of the *physical nature* of these 'states' – or indeed, of the physical nature of the whole machine. (Shall it consist of electronic relays, of cardboard, of human clerks sitting at desks, or what?) In other words, a given 'Turing machine' is an *abstract* machine which may be physically realized in an almost infinite number of different ways.

As soon as a Turing machine is physically realized, however, something interesting happens. Although the machine has from the logician's point of view only the states *A*, *B*, *C*, etc., it has from the engineer's point of view an almost infinite number of additional 'states' (though not in the same sense of 'state' – we shall call these *structural states*). For instance, if the machine consists of vacuum tubes, one of the things that may happen is that one of its vacuum tubes may fail – this puts the machine in what is from the physicist's if not the logician's point of view a different 'state'. Again, if the machine is a manually operated one built of cardboard, one of its possible 'non-logical' or 'structural' states is obviously that its cardboard may buckle. And so on.

A physically realized Turing machine may have no way of ascertaining its own structural state, just as a human being may have no way of ascertaining the condition of his appendix at a given time. However, it is extremely convenient to give a machine electronic 'sense organs' which enable it to scan itself and to detect minor malfunctions. These 'sense organs' may be visualized as causing certain symbols to be printed on an 'input tape' which the machine 'examines' from time to time. (One minor difficulty is that the 'report' of a sense organ might occupy a number of squares of tape, whereas the machine only 'scans'

one square at a time – however, this is unimportant, since it is well known that the effect of 'reading' any finite number of squares can be obtained using a program which only requires one square to be scanned at a time.)

(By way of a digression, let me remark that the first actually constructed digital computers did not have any devices of the kind just envisaged. On the other hand, they *did* have over 3000 vacuum tubes, some of which were failing at any given time! The need for 'routines' for self-checking therefore quickly became evident.)†

A machine which is able to detect at least some of its own structural states is in a position very analogous to that of a human being, who can detect some but not all of the malfunctions of his own body, and with varying degrees of reliability. Thus, suppose the machine 'prints out': 'Vacuum tube 312 has failed'. The question 'How did the machine ascertain that vacuum tube 312 failed?' is a perfectly sensible question. And the answer may involve a reference to both the physical structure of the machine ('sense organs', etc.) and the 'logical structure' (program for 'reading' and 'interpreting' the input tape).

If the machine prints: 'Vacuum tube 312 has failed' when vacuum tube 312 is in fact functioning, the mistake may be due to a miscomputation (in the course of 'reading' and 'interpreting' the input tape) or to an incorrect signal from a sense organ. On the other hand, if the machine prints: 'I am in state A', and it does this simply because its machine table contains the instruction: *Print: 'I am in state A when in state A'*, then the question of a miscomputation cannot arise. Even if some accident causes the printing mechanism to print: 'I am in state A' when the machine is *not* in state A, there was not a 'miscomputation' (only, so to speak, a 'verbal slip').

It is interesting to note that just as there are two possible descriptions of the behavior of a Turing machine – the engineer's structural blueprint and the logician's 'machine table' – so there are two possible descriptions of human psychology. The 'behavioristic' approach (including in this category theories which employ 'hypothetical constructs', including 'constructs' taken from physiology) aims at eventually providing a complete physicalistic‡ description of human behavior, in terms which link up with chemistry and physics. This corresponds to the engineer's or physicist's description of a physically realized Turing

† Actually, it was not necessary to add any 'sense organs'; existing computers check themselves by 'performing crucial experiments with themselves' (i.e. carrying out certain test computations and comparing the results with the correct results which have been given).

‡ In the sense of Oppenheim, 1958; not in the 'epistemological' sense associated with Carnap's writing on 'physicalism'.

machine. But it would also be possible to seek a more abstract description of human mental processes, in terms of 'mental states' (physical realization, if any, unspecified) and 'impressions' (these play the role of symbols on the machine's tapes) – a description which would specify the laws controlling the order in which the states succeeded one another, and the relation to verbalization (or, at any rate, verbalized thought). This description, which would be the analogue of a 'machine table', it was in fact the program of classical psychology to provide! Classical psychology is often thought to have failed for *methodological* reasons; I would suggest, in the light of this analogy, that it failed rather for empirical reasons – the mental states and 'impressions' of human beings do not orm a causally closed system to the extent to which the 'configurations' of a Turing machine do.

The analogy which has been presented between logical states of a Turing machine and mental states of a human being, on the one hand, and structural states of a Turing machine and physical states of a human being, on the other, is one that I find very suggestive. In particular, further exploration of this analogy may make it possible to further clarify the notion of a 'mental state' that we have been discussing. This 'further exploration' has not yet been undertaken, at any rate by me, but I should like to put down, for those who may be interested, a few of the features that seem to distinguish logical and mental states respectively from structural and physical ones:

(1) The functional organization (problem solving, thinking) of the human being or machine can be described in terms of the sequences of mental or logical states respectively (and the accompanying verbalizations), without reference to the nature of the 'physical realization' of these states.

(2) The states seem intimately connected with *verbalization*.

(3) In the case of rational thought (or computing), the 'program' which determines which states follow which, etc., is open to rational criticism.

4. Mind–body 'identity'

The last area in which we have to compare human beings and machines involves the question of *identifying* mental states with the corresponding physical states (or logical states with the corresponding structural states). As indicated at the beginning of this paper, all of the arguments for and against such identification can perfectly well be discussed in terms of Turing machines.

For example, in the 1930s Wittgenstein used the following argument: if I observe an after-image, and observe at the same time my brain state (with the aid of a suitable instrument) I observe *two* things, not one. (Presumably this is an argument *against* identification.) But we can perfectly well imagine a 'clever' Turing machine 'reasoning' as follows: 'When I print "I am in state A" I do not have to use my "sense organs". When I do use my "sense organs", and compare the occasions upon which I am in state A with the occasions upon which flip-flop 36 is on, I am comparing *two* things and not one.' And I do not think that we would find the argument of this mechanical Wittgenstein very convincing!

By contrast, Russell once carried the 'identity' view to the absurd extreme of maintaining that all we ever *see* is portions of our own brains. Analogously, a mechanical Russell might 'argue' that 'all I ever observe is my own vacuum tubes'. Both 'Russells' are wrong – the human being observes events in the outside world, and the process of 'observation' involves events in his brain. But we are not therefore forced to say that he 'really' observes his brain. Similarly, the machine T may 'observe', say, cans of tomato soup (if the machine's job is sorting cans of soup), and the process of 'observation' involves the functioning of vacuum tubes. But we are not forced to say that the machine 'really' observes its own vacuum tubes.

But let us consider more serious arguments on this topic. At the beginning of this paper, I pointed out that the *synthetic* character of the statement (1) 'I am in pain if, and only if, my C-fibers are stimulated' has been used as an argument for the view that the 'properties' (or 'events' or 'states') 'having C-fibers stimulated' and 'being in pain' cannot be the same. There are at least two reasons why this is not a very good argument: (a) the 'analytic-synthetic' distinction is not as sharp as that, especially where scientific laws are concerned; and (b) the criterion employed here for identifying 'properties' (or 'events' or 'states') is a very questionable one.

With respect to point (a): I have argued in chapter 2 that fundamental scientific laws cannot be happily classified as either 'analytic' or 'synthetic'. Consider, for example, the kind of conceptual shift that was involved in the transition from Euclidean to non-Euclidean geometry, or that would be involved if the law of the conservation of energy were to be abandoned. It is a distortion to say that the laws of Euclidean geometry (during their tenure of office) were 'analytic', and that Einstein merely 'changed the meaning of the words'. Indeed, it was precisely because Einstein did *not* change the meaning of the words, because he was really talking about shortest paths in the space in which

we live and move and have our being, that General Relativity seemed so incomprehensible when it was first proposed. To be told that one could come back to the same place by moving in one direction on a straight line! Adopting General Relativity was indeed adopting a whole new system of concepts – but that is not to say 'adopting a new system of verbal labels'.

But if it is a distortion to assimilate the revision of fundamental scientific laws to the adoption of new linguistic conventions, it is equally a mistake to follow conventional philosophers of science, and assimilate the conceptual change that Einstein inaugurated to the kind of change that arises when we discover a black swan (whereas we had previously assumed all swans to be white)! Fundamental laws are like principles of pure mathematics (as Quine has emphasized), in that they cannot be overthrown by isolated experiments: we can always hold on to the laws, and explain the experiments in various more or less *ad hoc* ways. And – in spite of the pejorative flavor of '*ad hoc*' – it is even *rational* to do this, in the case of important scientific theories, *as long as no acceptable alternative theory exists*. This is why it took a century of concept formation – and not just some experiments – to overthrow Euclidean geometry. And similarly, this is why we cannot today describe *any* experiments which would *by themselves* overthrow the law of the conservation of energy – although that law is not 'analytic', and might be abandoned if a new Einstein were to suggest good *theoretical* reasons for abandoning it, plus supporting experiments.

As Hanson has put it (Hanson, 1958), our concepts have theories 'built into' them – thus, to abandon a major scientific theory without providing an alternative would be to 'let our concepts crumble'. By contrast, although we *could* have held on to 'all swans are white' in the face of conflicting evidence, there would have been no *point* in doing so – the concepts involved did not *rest* on the acceptance of this or some rival principle in the way that geometrical concepts rest on the acceptance, not necessarily of Euclidean geometry, but of *some* geometry.

I do not deny that *today* any newly-discovered 'correlation' of the form: 'One is in mental state ψ if, and only if, one is in brain state ϕ' would *at first* be a *mere* correlation, a pure 'empirical generalization'. But I maintain that the interesting case is the case that would arise if we had a worked out and theoretically elaborated *system* of such 'correlations'. In such a case, scientific talk would be very different. Scientists would begin to say: 'It is impossible *in principle* to be in mental state ψ without being in brain state ϕ.' And it could very well be that the 'impossibility in principle' would amount to what Hanson rightly calls a

conceptual (Cf. Hanson, 1958) impossibility: scientists could not *conceive* (barring a new Einstein) of someone's being in mental state ψ without being in brain state ϕ. In particular, no experiment could *by itself* overthrow psychophysical laws which had acquired this kind of status.† Is it clear that in this kind of scientific situation it would not be correct to say that ϕ and ψ are the *same* state?

Moreover, the criteria for identifying 'events' or 'states' or 'properties' are by no means so clear. An example of a law with the sort of status we have been discussing is the following: Light passes through an aperture if, and only if, electromagnetic radiation (of such-and-such wavelengths) passes through the aperture.

This law is quite clearly *not* an 'analytic' statement. Yet it would be perfectly good scientific parlance to say that: (i) light passing through an aperture and (ii) electromagnetic radiation (of such-and-such wavelengths) passing through an aperture are two descriptions of the same event. (Indeed, in 'ordinary language' not only are descriptions of the same event not required to be equivalent: one may even speak of *incompatible* descriptions of the same event!)

It might be held, however, that *properties* (as opposed to events) cannot be described by different nonequivalent descriptions. Indeed, Frege, Lewis, and Carnap have *identified* properties and 'meanings' (so that *by definition* if two expressions have different meanings then they 'signify' different properties). This seems to me very dubious. But suppose it were correct. What would follow? One would have to admit that, e.g. being in pain and having C-fibers stimulated were different properties. But, in the language of the 'theory-constructing' Turing machine described at the beginning of this paper, one would equally have to admit that 'being in state A' and 'having flip-flop 36 on' were different properties. Indeed the sentences (i) 'I am in state A' and (ii) 'Flip-flop 36 is on' are clearly non-synonymous in the machine's language by any test (they have different syntactical properties and also different 'conditions of utterance' – e.g. the machine has to use different 'methods of verification'). Anyone who wishes, then, to argue on this basis for the existence of the soul will have to be prepared to hug the souls of Turing machines to his philosophical bosom!

5. A 'linguistic' argument

The last argument I shall consider on the subject of mind–body identity is a widely used 'linguistic' argument – it was, for example,

† Cf. the discussion of geometry in chapter 2 in this volume.

used by Max Black against Herbert Feigl at the Conference which inspired this volume (*Dimensions of Mind*). Consider the sentence:

(1) Pain *is identical with* stimulation of C-fibers.

The sentence is deviant (so the argument runs, though not in this terminology): there is no statement that it could be used to make in a normal context. Therefore, if a philosopher advances it as a thesis he must be giving the words a new meaning, rather than expressing any sort of discovery. For example (Max Black argued) one might begin to say 'I have stimulated C-fibers' instead of 'I have a pain', etc. But then one would *merely* be giving the expression 'has stimulated C-fibers' the new meaning 'is in pain'. The contention is that as long as the words keep their present meanings, (1) is unintelligible.

I agree that the sentence (1) is a 'deviant' sentence in present-day English. I do *not* agree that (1) can never become a normal, non-deviant sentence unless the words change their present meanings.

The point, in a nutshell, is that what is 'deviant' depends very much upon context, including the state of our knowledge, and with the development of new scientific theories it is constantly occurring that sentences that did not previously 'have a use', that were previously 'deviant', acquire a use – not because the words acquire *new* meanings, but because the old meanings as fixed by the core of stock uses, *determine* a new use given the new context.

There is nothing wrong with trying to bring linguistic theory to bear on this issue, but one must have a sufficiently sophisticated linguistic theory to bring to bear. The real question is not a question in *synchronic* linguistics but one in *diachronic*† linguistics, not 'Is (1) *now* a deviant sentence?' but 'If a change in scientific knowledge (e.g. the development of an integrated network of psychophysical laws of high "priority" in our overall scientific world view) were to lead to (1)'s becoming a *non*-deviant sentence, would a change in the meaning of a word necessarily have taken place?' – and this is not so simple a question.

Although this is not the time or the place to attempt the job of elaborating a semantical theory,‡ I should like to risk a few remarks on this question.

In the first place, it is easy to show that the mere uttering of a sentence

† Diachronic linguistics studies the language as it changes through time; synchronic linguistic seeks only to describe the language at one particular time.

‡ For a detailed discussion, cf. Ziff, 1960. I am extremely indebted to Ziff, both for making this work available to me and for personal communications on these matters. Section 5 of the present paper represents partly Ziff's influence (especially the use of the 'synchronic–diachronic' distinction), and partly the application of some of the ideas of chapter 2 in this volume to the present topic.

which no one has ever uttered before does not necessarily constitute the introduction of a 'new use'. If I say 'There is a purple Gila monster on this desk', I am very likely uttering a sentence that no English-speaker has uttered before me: but I am not in any way changing the meaning of any word.

In the second place, even if a sentence which was formerly deviant begins to acquire a standard use, no change in the *meaning* of any word need have taken place. Thus the sentence 'I am a thousand miles away from you', or its translation into ancient Greek, was undoubtedly a deviant sentence prior to the invention of *writing*, but acquired (was not 'given' but *acquired*) a normal use with the invention of writing and the ensuing possibility of long-distance interpersonal address.

Note the reasons that we would not say that any word (e.g. 'I', 'you', 'thousand') in this sentence changed its meaning: (a) the new use was not *arbitrary*, was not the product of *stipulation*, but represented an automatic projection† from the existing stock uses of the several words making up the sentence, given the new context; (b) the meaning of a sentence is in general a function of the meanings of the individual words making it up. (In fact this principle underlies the whole notion of word meaning – thus, if we said that the sentence had changed its meaning, we should have to face the question '*Which word* changed its meaning?'. But this would pretty clearly be an embarrassing question in this case.)

The case just described was one in which the new context was the product of new technology, but new theoretical knowledge may have a similar impact on the language. (For example, 'he went all the way around the world' would be a deviant sentence in a culture which did not know that the earth was round!) A case of this kind was discussed by Malcolm: we are beginning to have the means available for telling, on the basis of various physiological indicators (electroencephalograms, eye movements during sleep, blood pressure disturbances, etc.), when dreams begin and end. The sentence 'He is halfway through his dream' may, therefore, someday acquire a standard use. Malcolm's comment on this was that the words would in that case have been given a use. Malcolm is clearly mistaken, I believe; this case, in which a sentence acquires a use *because* of what the words mean is poles apart from the case in which words are literally *given* a use (i.e. in which meanings are stipulated for expressions). The 'realistic' account of this case is, I think, obviously correct: the sentence did not previously have a use because we had no way of telling when dreams start and stop. Now we are beginning to have ways of telling, and so we are beginning to find occasions upon which it is natural to employ this sentence. (Note that in

† The term is taken from Ziff, 1960.

Malcom's account there is no explanation of the fact that we give *this* sentence *this* use.)

Now, someone may grant that change in meaning should not be confused with change in distribution,† and that scientific and technological advances frequently produce changes in the latter that are not properly regarded as changes in the former. But one might argue that whereas one could have envisaged beforehand the circumstances under which the sentence 'He went all the way around the world.' would become nondeviant, one cannot now envisage any circumstances under which‡ 'Mental state ψ is identical with brain state ϕ.' would be nondeviant. But this is not a very good objection. In the first place, it might very well have been impossible for primitive people to envisage a spherical earth (the people on the 'underside' would obviously fall off). Even forty years ago, it might have been difficult if not impossible to envisage circumstances under which 'he is halfway through his dream' would be nondeviant. And in the second place, I believe that one *can* describe in general terms circumstances under which 'Mental state ψ is identical with brain state ϕ.' would become nondeviant.

In order to do this, it is necessary to talk about one important kind of 'is' – the *'is' of theoretical identification.* The use of 'is' in question is exemplified in the following sentences:

(2) Light is electromagnetic radiation (of such-and-such wavelengths).

(3) Water is H_2O.

What was involved in the scientific acceptance of, for instance, (2) was very roughly this: prior to the identification there were two distinct bodies of theory – optical theory (whose character Toulmin has very well described in his book on philosophy of science), and electromagnetic theory (as represented by Maxwell's equations). The decision to *define* light as 'electromagnetic radiation of such-and-such wavelengths' was scientifically justified by the following sorts of considerations (as has often been pointed out):

(1) It made possible the *derivation* of the laws of optics (up to first approximation) from more 'basic' physical laws. Thus, even if it had accomplished nothing else, this theoretical identification would have been a move towards simplifying the structure of scientific laws.

(2) It made possible the derivation of *new* predictions in the 'reduced' discipline (i.e. optics). In particular, it was now possible to predict that in certain cases the laws of geometrical optics would *not* hold. (Cf.

† The *distribution* of a word = the set of sentences in which it occurs.

‡ Here 'Mental state ψ is identical with brain state ϕ.' is used as a surrogate for such sentences as 'Pain is identical with stimulation of C-fibers.'

Duhem's famous comments on the reduction of Kepler's laws to Newton's.)

Now let us try to envisage the circumstances under which a theoretical identification of mental states with physiological states might be in accordance with good scientific procedure. In general terms, what is necessary is that we should have not *mere* 'correlates' for subjective states, but something much more elaborate – e.g. that we should know of physical states (say micro-states of the central processes) on the basis of which we could not merely *predict* human behaviour, but causally explain it.

In order to avoid 'category mistakes', it is necessary to restrict this notion, 'explain human behavior', very carefully. Suppose a man says 'I feel bad'. His behavior, described in one set of categories, is: 'stating that he feels bad'. And the explanation may be 'He said that he felt bad because he was hungry and had a headache'. I do not wish to suggest that the event 'Jones *stating* that he feels bad' can be explained in terms of the laws of *physics*. But there is *another* event which is very relevant, namely 'Jones's body producing such-and-such sound waves'. From one point of view this is a 'different event' from Jones's stating that he feels bad. But (to adapt a remark of Hanson's) there would be no point in remarking that these are different events if there were not a sense in which they were the *same* event. And it is the sense in which these are the 'same event' and not the sense in which these are 'different events' that is relevant here.

In fine, all I mean when I speak of 'causally explaining human behavior' is: causally explaining certain physical events (motions of bodies, productions of sound waves, etc.) which are in the sense just referred to the 'same' as the events which make up human behavior. And no amount of 'Ryle-ism' can succeed in arguing away† what is obviously a possibility: that physical science might succeed in doing this much.

If this much were a reality, then theoretically identifying 'mental states' with their 'correlates' would have the following two advantages:

(1) It would be possible (again up to 'first approximation') to derive from physical theory the classical laws (or low-level generalizations) of common-sense 'mentalistic' psychology, such as: 'People tend to avoid things with which they have had painful experiences'.

(2) It would be possible to predict the cases (and they are legion) in which common-sense 'mentalistic' psychology fails.

Advantage (2) could, of course, be obtained without 'identification'

† As one young philosopher attempted to do in a recent article in the *British Journal for the Philosophy of Science.*

(by using correlation laws). But advantage (2) could equally have been obtained in the case of optics without identification (by assuming that light *accompanies* electromagnetic radiation, but is not *identical* with it.) But the *combined* effect of eliminating certain laws altogether (in favor of theoretical definitions) *and* increasing the explanatory power of the theory could not be obtained in any other way in either case. The point worth noticing is that *every* argument for *and against* identification would apply equally in the mind–body case and in the light-electromagnetism case. (Even the 'ordinary language' argument could have been advanced against the identification of light with electromagnetic radiation.)

Two small points: (i) When I call 'light is electromagnetic radiation (of such-and-such wavelengths)' a definition, I do not mean that the statement is 'analytic'. But then 'definitions', *properly so called*, in theoretical science virtually *never* are analytic.† (Quine remarked once that he could think of at least nine good senses of 'definition' none of which had anything to do with analyticity.) Of course a philosopher might then object to the whole *rationale* of theoretical identification on the ground that it is no gain to eliminate 'laws' in favor of 'definitions' if both are *synthetic* statements. The fact that the scientist does not feel at all the same way is another illustration of how unhelpful it is to look at science from the standpoint of the question 'Analytic or synthetic?' (ii) Accepting a theoretical identification, e.g. 'Pain *is* stimulation of C-fibers', does not commit one to *interchanging* the terms 'pain' and 'stimulation of C-fibers' in idiomatic talk, as Black suggested. For instance, the identification of 'water' with 'H_2O' is by now a very well-known one, but no one says 'Bring me a glass of H_2O', except as a joke.

I believe that the account just presented is able (a) to explain the fact that sentences such as 'Mental state ψ is identical with brain state ϕ.' are deviant in present-day English, while (b) making it clear how these same sentences might become *non* deviant given a suitable increase in our scientific insight into the physical nature and causes of human behavior. The sentences in question cannot today be used to express a theoretical identification, because no such identification has been made. The act of theoretical identification is not an act that can be performed 'at will'; there are *preconditions* for its performance, as there are for many acts, and these preconditions are not satisfied today. On the other hand, if the sort of scientific theory described above should materialize, then the preconditions for theoretical identification would be met, as they were met in the light-electromagnetism case, and sentences of the type in

† This is argued in chapter 2.

question would then *automatically* require a use – namely, to express the appropriate theoretical identifications. Once again, what makes this way of *acquiring* a use different from being *given* a use (and from 'change of meaning' properly so called) is that the 'new use' is an automatic *projection* from existing uses, and does not involve arbitrary stipulation (except insofar as some element of 'stipulation' may be present in the acceptance of *any* scientific hypothesis, including 'The earth is round.').

So far we have considered only sentences of the form† 'Mental state ψ is identical with brain state ϕ.'. But what of the sentence:

(3) Mental states are micro-states of the brain.

This sentence does not, so to speak, 'give' any *particular* theoretical identification: it only says that unspecified theoretical identifications are possible. This is the sort of assertion that Feigl might make. And Black‡ might reply that in uttering (3) Feigl had uttered an odd set of words (i.e. a deviant sentence). It is possible that Black is right. Perhaps (3) is deviant in present-day English. But it is also possible that our descendants in two or three hundred years will feel that Feigl was making perfectly good sense and that the linguistic objections to (3) were quite silly. And they too may be right.

6. Machine linguistics

Let us consider the linguistic question that we have just discussed from the standpoint of the analogy between man and Turing machine that we have been presenting in this paper. It will be seen that our Turing machine will probably not be able, if it lacks suitable 'sense organs', to construct a correct theory of its own constitution. On the other hand 'I am in state A.' will be a sentence with a definite pattern of occurrence in the machine's 'language'. If the machine's 'language' is sufficiently complex, it may be possible to analyze it syntactically in terms of a finite set of basic building blocks (morphemes) and rules for constructing a potentially infinite set of 'sentences' from these. In particular, it will be possible to distinguish *grammatical*§ from *ungrammatical sentences* in the machine's 'language'. Similarly, it may be possible to associate regularities with sentence occurrences (or, 'describe sentence uses', in the Oxford jargon), and to assign 'meanings' to the finite set of morphemes and the finite set of forms of composition, in such a way that the 'uses'

† By sentences of this *form* I do not literally mean *substitution instances* of 'mental state ψ is identical with brain state ϕ.' Cf. note ‡ on page 379.

‡ I have, with hesitation, ascribed this position to Black on the basis of his remarks at the Conference. But, of course, I realize that he cannot justly be held responsible for remarks made on the spur of the moment.

§ This term is used in the sense of Chomsky, 1957, not in the traditional sense.

of the various sentences can be effectively projected from the meanings of the individual morphemes and forms of composition. In this case, one could distinguish not only 'grammatical' and 'ungrammatical' sentences in the 'machine language', but also 'deviant' and 'non-deviant' ones.

Chisholm would insist that it is improper to speak of machines as employing a language, and I agree. This is the reason for my occasionally enclosing the words 'language', 'meaning', etc., in 'raised-eyebrow' quotes – to emphasize, where necessary, that these words are being used in an extended sense. On the other hand, it is important to recognize that machine performances may be wholly *analogous* to language, so much so that the whole of linguistic theory can be applied to them. If the reader wishes to check this, he may go through a work like Chomsky's *Syntactic Structures* carefully, and note that *at no place is the assumption employed that the corpus of utterances studied by the linguist was produced by a conscious organism*. Then he may turn to such pioneer work in empirical semantics as Ziff's *Semantical Analysis* and observe that the same thing holds true for *semantical* theory.

Two further remarks in this connection: (i) Since I am contending that the mind–body problem is *strictly analogous* to the problem of the relation between structural and logical states, not that the two problems are *identical*, a suitable *analogy* between machine 'language' and human language is all that is needed here. (ii) Chisholm might contend that a 'behavioristic' semantics of the kind attempted by Ziff (i.e. one that does not take 'intentionality' as a primitive notion) is impossible. But even if this were true, it would not be relevant. For if *any* semantical theory can fit human language, it has to be shown why a completely *analogous* theory would not fit the language of a suitable machine. For instance, if 'intentionality' plays a role as a primitive notion in a *scientific* explanation of human language, then a theoretical construct with similar *formal* relations to the corresponding 'observables' will have the *same* explanatory power in the case of machine 'language'.

Of course, the objection to 'behavioristic' linguistics might *really* be an objection to all attempts at *scientific* linguistics. But this possibility I feel justified in dismissing.

Now suppose we equip our 'theory-constructing' Turing machine with 'sense organs' so that it can obtain the empirical data necessary for the construction of a theory of its own nature.

Then it may introduce into its 'theoretical language' noun phrases that can be 'translated' by the English expression 'flip-flop 36', and sentences that can be translated by 'Flip-flop 36 is on.'. These expressions will have a meaning and use quite distinct from the meaning and use of 'I am in state A.' in the machine language.

If any 'linguistic' argument really shows that the sentence 'Pain is identical with stimulation of C-fibers.' is deviant, in English, the same argument must show that 'State *A* is identical with flip-flop 36 being on' is deviant in the machine language. If any argument shows that 'Pain is identical with stimulation of C-fibers.' could not become non-deviant (viewing English now *dia*chronically) unless the words first altered their meanings, the same argument, applied to the 'diachronic linguistics of machine language', would show that the sentence 'State *A* is identical with flip-flop 36 being on.' could not become non-deviant in machine language unless the words first changed their meanings. In short, every philosophic argument that has ever been employed in connection with the mind–body problem, from the oldest and most naive (e.g. 'states of consciousness can just be *seen* to be different from physical states') to the most sophisticated, has its exact counterpart in the case of the 'problem' of logical states and structural states in Turing machines.

7. Conclusion

The moral, I believe, is quite clear: it is no longer possible to believe that the mind–body problem is a genuine theoretical problem, or that a 'solution' to it would shed the slightest light on the world in which we live. For it is quite clear that no grown man in his right mind would take the problem of the 'identity' or 'non-identity' of logical and structural states in a machine at all seriously – not because the answer is obvious, but because it is obviously of no importance *what* the answer is. But if the so-called 'mind–body problem' is nothing but a different realization of the same set of logical and linguistic issues, then it must be just as empty and just as verbal.

It is often an important insight that two problems with distinct subject matter are the same in all their logical and methodological aspects. In this case, the insight carries in its train the realization that any conclusion that might be reached in the case of the mind–body problem would have to be reached, *and for the same reasons*, in the Turing machine case. But if it is clear (as it obviously is) that, for example, the conclusion that the logical states of Turing machines are hopelessly different from their structural states, even if correct, could represent only a purely *verbal* discovery, then the same conclusion *reached by the same arguments* in the human case must likewise represent a purely verbal discovery. To put it differently, if the mind–body problem is identified with any problem of more than purely conceptual interest (e.g. with the question of whether or not human beings have 'souls')

then *either* it must be that (a) no argument *ever* used by a philosopher sheds the *slightest* light on it (and this independently of the way the argument tends), or (b) that some philosophic argument for mechanism is correct, or (c) that some dualistic argument does show that *both* human beings *and* Turing machines have souls! I leave it to the reader to decide which of the three alternatives is at all plausible.