Chapter 10

The Nature of Mental States

Hilary Putnam

The typical concerns of the Philosopher of Mind might be represented by three questions: (1) How do we know that other people have pains? (2) Are pains brain states? (3) What is the analysis of the concept *pain*? I do not wish to discuss questions (1) and (3) in this chapter. I shall say something about question (2).

Identity Questions

'Is pain a brain state?' (Or, 'Is the property of having a pain at time t a brain state?')[1] It is impossible to discuss this question sensibly without saying something about the peculiar rules which have grown up in the course of the development of 'analytical philosophy'—rules which, far from leading to an end to all conceptual confusions, themselves represent considerable conceptual confusion. These rules—which are, of course, implicit rather than explicit in the practice of most analytical philosophers—are (1) that a statement of the form 'being A is being B' (e.g., 'being in pain is being in a certain brain state') can be *correct* only if it follows, in some sense, from the meaning of the terms A and B; and (2) that a statement of the form 'being A is being B' can be philosophically *informative* only if it is in some sense reductive (e.g., 'being in pain is having a certain unpleasant sensation' is not philosophically informative; 'being in pain is having a certain behavior disposition' is, if true, philosophically informative). These rules are excellent rules if we still believe that the program of reductive analysis (in the style of the 1930s) can be carried out; if we don't, then they turn analytical philosophy into a mug's game, at least so far as 'is' questions are concerned.

In this paper I shall use the term 'property' as a blanket term for such things as being in pain, being in a particular brain state, having a particular behavior disposition, and also for magnitudes such as temperature, etc.—i.e., for things which can naturally be represented by one-or-more-place predicates or functors. I shall use the term 'concept' for things which can be identified with synonymy-classes of expressions. Thus the concept *temperature* can be identified (I maintain) with the synonymy-class of the word 'temperature'.[2] (This is like saying that the number 2 can be identified with the class of all pairs. This is quite a different statement from the peculiar statement that 2 *is* the class of all pairs. I do not maintain that concepts *are* synonymy-classes, whatever that might mean, but that they can be identified with synonymy-classes, for the purpose of formalization of the relevant discourse.)

The question 'What is the concept *temperature*?' is a very 'funny' one. One might take it to mean 'What is temperature? Please take my question as a conceptual one.' In that case an answer might be (pretend for a moment 'heat' and 'temperature' are synonyms) 'temperature is heat', or even 'the concept of temperature is the same concept as the concept of heat'. Or one might take it to mean 'What are *concepts*, really? For example, what is "the concept of temperature"?' In that case heaven knows what an 'answer'

would be. (Perhaps it would be the statement that concepts *can be identified with* synonymy-classes.)

Of course, the question 'What is the property temperature?' is also 'funny'. And one way of interpreting it is to take it as a question about the concept of temperature. But this is not the way a physicist would take it.

The effect of saying that the property P_1 can be identical with the property P_2 only if the terms P_1, P_2 are in some suitable sense 'synonyms' is, to all intents and purposes, to collapse the two notions of 'property' and 'concept' into a single notion. The view that concepts (intensions) *are* the same as properties has been explicitly advocated by Carnap (e.g., in *Meaning and Necessity*). This seems an unfortunate view, since 'temperature is mean molecular kinetic energy' appears to be a perfectly good example of a true statement of identity of properties, whereas 'the concept of temperature is the same concept as a concept of mean molecular kinetic energy' is simply false.

Many philosophers believe that the statement 'pain is a brain state' violates some rules or norms of English. But the arguments offered are hardly convincing. For example, if the fact that I can know that I am in pain without knowing that I am in brain state *S* shows that pain cannot be brain state *S*, then, by exactly the same argument, the fact that I can know that the stove is hot without knowing that the mean molecular kinetic energy is high (or even that molecules exist) shows that it is *false* that temperature is mean molecular kinetic energy, physics to the contrary. In fact, all that immediately follows from the fact that I can know that I am in pain without knowing that I am in brain state *S* is that the concept of pain is not the same concept as the concept of being in brain state *S*. But either pain, or the state of being in pain, or some pain, or some pain state, might still be brain state *S*. After all, the concept of temperature is not the same concept as the concept of mean molecular kinetic energy. But temperature is mean molecular kinetic energy.

Some philosophers maintain that both 'pain is a brain state' and 'pain states are brain states' are unintelligible. The answer is to explain to these philosophers, as well as we can, given the vagueness of all scientific methodology, what sorts of considerations lead one to make an empirical reduction (i.e., to say such things as 'water is H_2O', 'light is electromagnetic radiation', 'temperature is mean molecular kinetic energy'). If, without giving reasons, he still maintains in the face of such examples that one cannot imagine parallel circumstances for the use of 'pains are brain states' (or, perhaps, 'pain states are brain states'), one has grounds to regard him as perverse.

Some philosophers maintain that 'P_1 is P_2' is something that can be true, when the 'is' involved is the 'is' of empirical reduction, only when the properties P_1 and P_2 are (a) associated with a spatio-temporal region; and (b) the region is one and the same in both cases. Thus 'temperature is mean molecular kinetic energy' is an admissible empirical reduction, since the temperature and the molecular energy are associated with the same space-time region, but 'having a pain in my arm is being in a brain state' is not, since the spatial regions involved are different.

This argument does not appear very strong. Surely no one is going to be deterred from saying that mirror images are light reflected from an object and then from the surface of a mirror by the fact that an image can be 'located' three feet *behind* the mirror! (Moreover, one can always find *some* common property of the reductions one is willing to allow—e.g., temperature is mean molecular kinetic energy—which is not a property of some one identification one wishes to disallow. This is not very impressive unless one has an argument to show that the very purposes of such identification depend upon the common property in question.)

Again, other philosophers have contended that all the predictions that can be derived from the conjunction of neurophysiological laws with such statements as 'pain states are such-and-such brain states' can equally well be derived from the conjunction of the same neurophysiological laws with 'being in pain is correlated with such-and-such brain states', and hence (*sic*!) there can be no methodological grounds for saying that pains (or pain states) *are* brain states, as opposed to saying that they are *correlated* (invariantly) with brain states. This argument, too, would show that light is only correlated with electromagnetic radiation. The mistake is in ignoring the fact that, although the theories in question may indeed lead to the same predictions, they open and exclude different *questions*. 'Light is invariantly correlated with electromagnetic radiation' would leave open the questions 'What is the light then, if it isn't the same as the electromagnetic radiation?' and 'What makes the light accompany the electromagnetic radiation?'—questions which are excluded by saying that the light *is* the electromagnetic radiation. Similarly, the purpose of saying that pains are brain states is precisely to exclude from empirical meaningfulness the questions 'What is the pain, then, if it isn't the same as the brain state?' and 'What makes the pain accompany the brain state?' If there are grounds to suggest that these questions represent, so to speak, the wrong way to look at the matter, then those grounds are grounds for a theoretical identification of pains with brain states.

If all arguments to the contrary are unconvincing, shall we then conclude that it is meaningful (and perhaps true) to say either that pains are brain states or that pain states are brain states?

1. It is perfectly meaningful (violates no 'rule of English', involves no 'extension of usage') to say 'pains are brain states'.
2. It is not meaningful (involves a 'changing of meaning' or 'an extension of usage', etc.) to say 'pains are brain states'.

My own position is not expressed by either 1 or 2. It seems to me that the notions 'change of meaning' and 'extension of usage' are simply so ill defined that one cannot in fact say *either* 1 or 2. I see no reason to believe that either the linguist, or the man-on-the-street, or the philosopher possesses today a notion of 'change of meaning' applicable to such cases as the one we have been discussing. The *job* for which the notion of change of meaning was developed in the history of the language was just a *much* cruder job than this one.

But, if we don't assert either 1 or 2—in other words, if we regard the 'change of meaning' issue as a pseudo-issue in this case—then how are we to discuss the question with which we started? 'Is pain a brain state?'

The answer is to allow statements of the form 'pain is *A*', where 'pain' and '*A*' are in no sense synonyms, and to see whether any such statement can be found which might be acceptable on empirical and methodological grounds. This is what we shall now proceed to do.

Is Pain a Brain State?

We shall discuss 'Is pain a brain state?' then. And we have agreed to waive the 'change of meaning' issue.

Since I am discussing not what the concept of pain comes to, but what pain is, in a sense of 'is' which requires empirical theory-construction (or, at least, empirical speculation), I shall not apologize for advancing an empirical hypothesis. Indeed, my strategy

will be to argue that pain is *not* a brain state, not on *a priori* grounds, but on the grounds that another hypothesis is more plausible. The detailed development and verification of my hypothesis would be just as Utopian a task as the detailed development and verification of the brain-state hypothesis. But the putting-forward, not of detailed and scientifically 'finished' hypotheses, but of schemata for hypotheses, has long been a function of philosophy. I shall, in short, argue that pain is not a brain state, in the sense of a physical-chemical state of the brain (or even the whole nervous system), but another *kind* of state entirely. I propose the hypothesis that pain, or the state of being in pain, is a functional state of a whole organism.

To explain this it is necessary to introduce some technical notions. In previous papers I have explained the notion of a Turing Machine and discussed the use of this notion as a model for an organism. The notion of a Probabilistic Automaton is defined similarly to a Turing Machine, except that the transitions between 'states' are allowed to be with various probabilities rather than being 'deterministic'. (Of course, a Turing Machine is simply a special kind of Probabilistic Automaton, one with transition probabilities 0, 1). I shall assume the notion of a Probabilistic Automaton has been generalized to allow for 'sensory inputs' and 'motor outputs'—that is, the Machine Table specifies, for every possible combination of a 'state' and a complete set of 'sensory inputs', an 'instruction' which determines the probability of the next 'state', and also the probabilities of the 'motor outputs'. (This replaces the idea of the Machine as printing on a tape.) I shall also assume that the physical realization of the sense organs responsible for the various inputs, and of the motor organs, is specified, but that the 'states' and the 'inputs' themselves are, as usual, specified only 'implicitly'—i.e., by the set of transition probabilities given by the Machine Table.

Since an empirically given system can simultaneously be a 'physical realization' of many different Probabilistic Automata, I introduce the notion of a *Description* of a system. A Description of S where S is a system, is any true statement to the effect that S possesses distinct states $S_1, S_2 \ldots S_n$ which are related to one another and to the motor outputs and sensory inputs by the transition probabilities given in such-and-such a Machine Table. The Machine Table mentioned in the Description will then be called the Functional Organization of S relative to that Description, and the S_i such that S is in state S_i at a given time will be called the Total State of S (at the time) relative to that Description. It should be noted that knowing the Total State of a system relative to a Description involves knowing a good deal about how the system is likely to 'behave', given various combinations of sensory inputs, but does *not* involve knowing the physical realization of the S_i as, e.g., physical-chemical states of the brain. The S_i, to repeat, are specified only *implicitly* by the Description—i.e., specified *only* by the set of transition probabilities given in the Machine Table.

The hypothesis that 'being in pain is a functional state of the organism' may now be spelled out more exactly as follows:

1. All organisms capable of feeling pain are Probabilistic Automata.
2. Every organism capable of feeling pain possesses at least one Description of a certain kind (i.e., being capable of feeling pain *is* possessing an appropriate kind of Functional Organization).
3. No organism capable of feeling pain possesses a decomposition into parts which separately possess Descriptions of the kind referred to in 2.
4. For every Description of the kind referred to in 2, there exists a subset of the sensory inputs such that an organism with that Description is in pain when and only when some of its sensory inputs are in that subset.

This hypothesis is admittedly vague, though surely no vaguer than the brain-state hypothesis in its present form. For example, one would like to know more about the kind of Functional Organization that an organism must have to be capable of feeling pain, and more about the marks that distinguish the subset of the sensory inputs referred to in 4. With respect to the first question, one can probably say that the Functional Organization must include something that resembles a 'preference function', or at least a preference partial ordering and something that resembles an 'inductive logic' (i.e., the Machine must be able to 'learn from experience'). In addition, it seems natural to require that the Machine possess 'pain sensors', i.e., sensory organs which normally signal damage to the Machine's body, or dangerous temperatures, pressures, etc., which transmit a special subset of the inputs, the subset referred to in 4. Finally, and with respect to the second question, we would want to require at least that the inputs in the distinguished subset have a high disvalue on the Machine's preference function or ordering (further conditions are discussed in the previous chapter). The purpose of condition 3 is to rule out such 'organisms' (if they can count as such) as swarms of bees as single pain-feelers. The condition 1 is, obviously, redundant, and is only introduced for expository reasons. (It is, in fact, empty, since everything is a Probabilistic Automaton under *some* Description.)

I contend, in passing, that this hypothesis, in spite of its admitted vagueness, is far *less* vague than the 'physical-chemical state' hypothesis is today, and far more susceptible to investigation of both a mathematical and an empirical kind. Indeed, to investigate this hypothesis is just to attempt to produce 'mechanical' models of organisms—and isn't this, in a sense, just what psychology is about? The difficult step, of course, will be to pass from models of *specific* organisms to a *normal form* for the psychological description of organisms—for this is what is required to make 2 and 4 precise. But this too seems to be an inevitable part of the program of psychology.

I shall now compare the hypothesis just advanced with (a) the hypothesis that pain is a brain state, and (b) the hypothesis that pain is a behavior disposition.

Functional State versus Brain State

It may, perhaps, be asked if I am not somewhat unfair in taking the brain-state theorist to be talking about *physical-chemical* states of the brain. But (a) these are the only sorts of states ever mentioned by brain-state theorists. (b) The brain-state theorist usually mentions (with a certain pride, slightly reminiscent of the Village Atheist) the incompatibility of his hypothesis with all forms of dualism and mentalism. This is natural if physical-chemical states of the brain are what is at issue. However, functional states of whole systems are something quite different. In particular, the functional-state hypothesis is *not* incompatible with dualism! Although it goes without saying that the hypothesis is 'mechanistic' in its inspiration, it is a slightly remarkable fact that a system consisting of a body and a 'soul', if such things there be, can perfectly well be a Probabilistic Automaton. (c) One argument advanced by Smart is that the brain-state theory assumes only 'physical' properties, and Smart finds 'non-physical' properties unintelligible. The Total States and the 'inputs' defined above are, of course, neither mental nor physical *per se*, and I cannot imagine a functionalist advancing this argument. (d) If the brain-state theorist does mean (or at least allow) states other than physical-chemical states, then his hypothesis is completely empty, at least until he specifies *what* sort of 'states' he *does* mean.

Taking the brain-state hypothesis in this way, then, what reasons are there to prefer the functional-state hypothesis over the brain-state hypothesis? Consider what the

brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that *any* organism (not just a mammal) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusca, and certainly feel pain), etc. At the same time, it must *not* be a possible (physically possible) state of the brain of any physically possible creature that cannot feel pain. Even if such a state can be found, it must be nomologically certain that it will also be a state of the brain of any extraterrestrial life that may be found that will be capable of feeling pain before we can even entertain the supposition that it may *be* pain.

It is not altogether impossible that such a state will be found. Even though octopus and mammal are examples of parallel (rather than sequential) evolution, for example, virtually identical structures (physically speaking) have evolved in the eye of the octopus and in the eye of the mammal, notwithstanding the fact that this organ has evolved from different kinds of cells in the two cases. Thus it is at least possible that parallel evolution, all over the universe, might *always* lead to *one and the same* physical 'correlate' of pain. But this is certainly an ambitious hypothesis.

Finally, the hypothesis becomes still more ambitious when we realize that the brain-state theorist is not just saying that *pain* is a brain state; he is, of course, concerned to maintain that *every* psychological state is a brain state. Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say 'hungry'), but whose physical-chemical 'correlate' is different in the two cases, the brain-state theory has collapsed. It seems to me overwhelmingly probable that we can do this. Granted, in such a case the brain-state theorist can save himself by *ad hoc* assumptions (e.g., defining the disjunction of two states to be a single 'physical-chemical state'), but this does not have to be taken seriously.

Turning now to the considerations *for* the functional-state theory, let us begin with the fact that we identify organisms as in pain, or hungry, or angry, or in heat, etc., on the basis of their *behavior*. But it is a truism that similarities in the behavior of two systems are at least a reason to suspect similarities in the functional organization of the two systems, and a much *weaker* reason to suspect similarities in the actual physical details. Moreover, we expect the various psychological states—at least the basic ones, such as hunger, thirst, aggression, etc.—to have more or less similar 'transition probabilities' (within wide and ill-defined limits, to be sure) with each other and with behavior in the case of different species, because this is an artifact of the way in which we identify these states. Thus, we would not count an animal as *thirsty* if its 'unsatiated' behavior did not seem to be directed toward drinking and was not followed by 'satiation for liquid'. Thus any animal that we count as capable of these various states will at least *seem* to have a certain rough kind of functional organization. And, as already remarked, if the program of finding psychological laws that are not species-specific—i.e., of finding a normal form for psychological theories of different species—ever succeeds, then it will bring in its wake a delineation of the kind of functional organization that is necessary and sufficient for a given psychological state, as well as a precise definition of the notion 'psychological state'. In contrast, the brain-state theorist has to hope for the eventual development of neurophysiological laws that are species-independent, which seems much less reasonable than the hope that psychological laws (of a sufficiently general kind) may be species-independent, or, still weaker, that a species-independent *form* can be found in which psychological laws can be written.

Functional State versus Behavior-Disposition

The theory that being in pain is neither a brain state nor a functional state but a behavior disposition has one apparent advantage: it appears to agree with the way in which we verify that organisms are in pain. We do not in practice know anything about the brain state of an animal when we say that it is in pain; and we possess little if any knowledge of its functional organization, except in a crude intuitive way. In fact, however, this 'advantage' is no advantage at all: for, although statements about how we verify that x is A may have a good deal to do with what the concept of being A comes to, they have precious little to do with what the property A *is*. To argue on the ground just mentioned that pain is neither a brain state nor a functional state is like arguing that heat is not mean molecular kinetic energy from the fact that ordinary people do not (they think) ascertain the mean molecular kinetic energy of something when they verify that it is hot or cold. It is not necessary that they should; what is necessary is that the marks that they take as indications of heat should in fact be explained by the mean molecular kinetic energy. And, similarly, it is necessary to our hypothesis that the marks that are taken as behavioral indications of pain should be explained by the fact that the organism is a functional state of the appropriate kind, but not that speakers should *know* that this is so.

The difficulties with 'behavior disposition' accounts are so well known that I shall do little more than recall them here. The difficulty—it appears to be more than a 'difficulty,' in fact—of specifying the required behavior disposition except as 'the disposition of X to behave as if X were in *pain*', is the chief one, of course. In contrast, we *can* specify the functional state with which we propose to identify pain, at least roughly, without using the notion of pain. Namely, the functional state we have in mind is the state of receiving sensory inputs which play a certain role in the Functional Organization of the organism. This role is characterized, at least partially, by the fact that the sense organs responsible for the inputs in question are organs whose function is to detect damage to the body, or dangerous extremes of temperature, pressure, etc., and by the fact that the 'inputs' themselves, whatever their physical realization, represent a condition that the organism assigns a high disvalue to. As I stressed in 'The mental life of some machines', this does *not* mean that the Machine will always *avoid* being in the condition in question ('pain'); it only means that the condition will be avoided unless not avoiding it is necessary to the attainment of some more highly valued goal. Since the behavior of the Machine (in this case, an organism) will depend not merely on the sensory inputs, but also on the Total State (i.e., on other values, beliefs, etc.), it seems hopeless to make any general statement about how an organism in such a condition *must* behave; but this does not mean that we must abandon hope of characterizing the condition. Indeed, we have just characterized it.

Not only does the behavior-disposition theory seem hopelessly vague; if the 'behavior' referred to is peripheral behavior, and the relevant stimuli are peripheral stimuli (e.g., we do not say anything about what the organism will do if its brain is operated upon), then the theory seems clearly false. For example, two animals with all motor nerves cut will have the same actual and potential 'behavior' (namely, none to speak of); but if one has cut pain fibers and the other has uncut pain fibers, then one will feel pain and the other won't. Again, if one person has cut pain fibers, and another suppresses all pain responses deliberately due to some strong compulsion, then the actual and potential peripheral behavior may be the same, but one will feel pain and the other won't. (Some philosophers maintain that this last case is conceptually impossible, but the only evidence for this appears to be that *they* can't, or don't want to, conceive of it.) If, instead of pain, we take some sensation the 'bodily expression' of which is easier

to suppress—say, a slight coolness in one's left little finger—the case becomes even clearer.

Finally, even if there *were* some behavior disposition invariantly correlated with pain (species-independently!), and specifiable without using the term 'pain', it would still be more plausible to identify being in pain with some state whose presence *explains* this behavior disposition—the brain state or functional state—than with the behavior disposition itself. Such considerations of plausibility may be somewhat subjective; but if other things *were* equal (of course, they aren't) why shouldn't we allow considerations of plausibility to play the deciding role?

Methodological Considerations

So far we have considered only what might be called the 'empirical' reasons for saying that being in pain is a functional state, rather than a brain state or a behavior disposition; namely, that it seems more likely that the functional state we described is invariantly 'correlated' with pain, species-independently, than that there is either a physical-chemical state of the brain (must an organism have a *brain* to feel pain? perhaps some ganglia will do) or a behavior disposition so correlated. If this is correct, then it follows that the identification we proposed is at least a candidate for consideration. What of methodological considerations?

The methodological considerations are roughly similar in all cases of reduction, so no surprises need be expected here. First, identification of psychological states with functional states means that the laws of psychology can be derived from statements of the form 'such-and-such organisms have such-and-such Descriptions' together with the identification statements ('being in pain is such-and-such a functional state', etc.). Secondly, the presence of the functional state (i.e., of inputs which play the role we have described in the Functional Organization of the organism) is not merely 'correlated with' but actually explains the pain behavior on the part of the organism. Thirdly, the identification serves to exclude questions which (if a naturalistic view is correct) represent an altogether wrong way of looking at the matter, e.g., 'What *is* pain if it isn't either the brain state or the functional state?' and 'What causes the pain to be always accompanied by this sort of functional state?' In short, the identification is to be tentatively accepted as a theory which leads to both fruitful predictions and to fruitful *questions*, and which serves to discourage fruitless and empirically senseless questions, where by 'empirically senseless' I mean 'senseless' not merely from the standpoint of verification, but from the standpoint of what there in fact *is*.

Notes

1. In this paper I wish to avoid the vexed question of the relation between *pains* and *pain states*. I only remark in passing that one common argument *against* identification of these two—namely, that a pain can be in one's arm but a state (of the organism) cannot be in one's arm—is easily seen to be fallacious.
2. There are some well-known remarks by Alonzo Church on this topic. Those remarks do not bear (as might at first be supposed) on the identification of concepts with synonymy-classes as such, but rather support the view that (in formal semantics) it is necessary to retain Frege's distinction between the normal and the 'oblique' use of expressions. That is, even if we say that the concept of temperature *is* the synonymy-class of the word 'temperature', we must not thereby be led into the error of supposing that 'the concept of temperature' is synonymous with 'the synonymy-class of the word "temperature"'—for then 'the concept of temperature' and *'der Begriff der Temperatur'* would not be synonymous, which they are. Rather, we must say that the concept of 'temperature' *refers to* the synonymy-class of the word 'temperature' (on this particular reconstruction); but that class is *identified* not as 'the synonymy-class to which such-and-such a word belongs', but in another way (e.g., as the synonymy-class whose members have such-and-such a characteristic use).