2

Brains and Behavior

Hilary Putnam

Once upon a time there was a tough-minded philosopher who said, 'What is all this talk about "minds", "ideas", and "sensations"? Really—and I mean *really* in the real world—there is nothing to these so-called "mental" events and entities but certain processes in our all-too-material heads.'

And once upon a time there was a philosopher who retorted, 'What a masterpiece of confusion! Even if, say, *pain* were perfectly correlated with any particular event in my brain (which I doubt) that event would obviously have certain properties—say, a certain numerical intensity measured in volts—which it would be *senseless* to ascribe to the feeling of pain. Thus, it is *two* things that are correlated, not *one*—and to call *two* things *one* thing is worse than being mistaken; it is utter contradiction.'

For a long time dualism and materialism appeared to exhaust the alternatives. Compromises were attempted ('double aspect' theories), but they never won many converts and practically no one found them intelligible. Then, in the mid-1930s, a seeming third possibility was discovered. This third possibility has been called *logical behaviorism*. To state the nature of this third possibility briefly, it is necessary to recall the treatment of the natural numbers (i.e. zero, one, two, three . . .) in modern logic. Numbers are identified with *sets*, in various ways, depending on which authority one follows. For instance, Whitehead and Russell identified zero with the set of all empty sets, one with the set of all one-membered sets, two with the set of all two-membered sets, three with the set of all three-membered sets, and so on. (This has the appearance of circularity, but they were able to dispel this appearance by defining 'one-membered set', 'two-membered set', 'three-membered set', etc., without using 'one', 'two', 'three', etc.) In short, numbers are treated as *logical constructions out of sets*. The number theorist is doing set theory without knowing it, according to this interpretation.

What was novel about this was the idea of getting rid of certain philosophi-

From R. J. Butler, ed., *Analytical Philosophy*, vol. 2 (Oxford: Blackwell, 1965). Reprinted by permission of the author. Notes have been renumbered for this edition. This paper was read as part of the program of the American Association for the Advancement of Science, Section L (History and Philosophy of Science), December 27, 1961.

cally unwanted or embarrassing entities (numbers) without failing to do justice to the appropriate body of discourse (number theory) by treating the entities in question as logical constructions. Russell was quick to hold up this success as a model to all future philosophers. And certain of those future philosophers—the Vienna positivists, in their 'physicalist' phase (about 1930)—took Russell's advice so seriously as to produce the doctrine that we are calling *logical behaviorism*—the doctrine that, just as numbers are (allegedly) logical constructions out of *sets*, so *mental events* are logical constructions out of actual and possible *behavior events*.

In the set theoretic case, the 'reduction' of number theory to the appropriate part of set theory was carried out in detail and with indisputable technical success. One may dispute the philosophical significance of the reduction, but one knows exactly what one is talking about when one disputes it. In the mind-body case, the reduction was never carried out in even *one* possible way, so that it is not possible to be clear on just *how* mental entities or events are to be (identified with) logical constructions out of behavior events. But broadly speaking, it is clear what the view implies: it implies that all talk about mental events is translatable into talk about actual or potential overt behavior.

It is easy to see in what way this view differs from both dualism and classical materialism. The logical behaviorist agrees with the dualist that what goes on in our brains has no connection whatsoever with what we *mean* when we say that someone is in pain. He can even take over the dualist's entire stock of arguments against the materialist position. Yet, at the same time, he can be as 'tough-minded' as the materialist in denying that ordinary talk of 'pains', 'thoughts', and 'feelings' involves reference to 'Mind' as a Cartesian substance.

Thus it is not surprising that logical behaviorism attracted enormous attention —both pro and con—during the next thirty years. Without doubt, this alternative proved to be a fruitful one to inject into the debate. Here, however, my intention is not to talk about the fruitfulness of the investigations to which logical behaviorism has led, but to see if there was any upshot to those investigations. Can we, after thirty years, say anything about the rightness or wrongness of logical behaviorism? Or must we say that a third alternative has been added to the old two; that we cannot decide between three any more easily than we could decide between two; and that our discussion is thus half as difficult again as it was before?

One conclusion emerged very quickly from the discussion pro and con logical behaviorism: that the extreme thesis of logical behaviorism, as we just stated it (that all talk about 'mental events' is translatable into talk about overt behavior) is false. But, in a sense, this is not very interesting. An extreme thesis may be false, although there is 'something to' the way of thinking that it represents. And the more interesting question is this: what, if anything, can be 'saved' of the way of thinking that logical behaviorism represents?

In the last thirty years, the original extreme thesis of logical behaviorism has gradually been weakened to something like this:

(1) That there exist entailments between mind-statements and behavior-statements; entailments that are not, perhaps, analytic in the way in which 'All bachelors are unmarried' is analytic, but that nevertheless follow (in some sense) from the meanings of mind words. I shall call these *analytic entailments*.

(2) That these entailments may not provide an actual *translation* of 'mind talk' into 'behavior talk' (this 'talk' talk was introduced by Gilbert Ryle in his *Concept of Mind*), but that this is true for such superficial reasons as the greater ambiguity of mind talk, as compared with the rela-

tively greater specificity of overt behavior talk.

I believe that, although no philosopher would to-day subscribe to the older version of behaviorism, a great many philosophers[1] would accept these two points, while admitting the unsatisfactory imprecision of the present statement of both of them. If these philosophers are right, then there is much work to be done (e.g. the notion of 'analyticity' has to be made clear), but the direction of work is laid out for us for some time to come.

I wish that I could share this happy point of view—if only for the comforting conclusion that first-rate philosophical research, continued for some time, will eventually lead to a solution to the mind-body problem which is independent of troublesome empirical facts about brains, central causation of behavior, evidence for and against nonphysical causation of at least some behavior, and the soundness or unsoundness of psychical research and parapsychology. But the fact is that I come to bury logical behaviorism, not to praise it. I feel that the time has come for us to admit that logical behaviorism is a mistake, and that even the weakened forms of the logical behaviorist doctrine are incorrect. I cannot hope to establish this in so short a paper as this one;[2] but I hope to expose for your inspection at least the main lines of my thinking.

Logical Behaviorism

The logical behaviorist usually begins by pointing out what is perfectly true, that such words as 'pain' ('pain' will henceforth be our stock example of a mind word) are not taught by reference to standard examples in the way in which such words as 'red' are. One can point to a standard red thing, but one cannot point to a standard pain (that is, except by pointing to some piece of *behavior*) and say: 'Compare the feeling you are having with this one (say, Jones's feeling at time t_1). If the two feelings have the identical *quality*, then your feeling is legitimately calle feeling of *pain.*' The difficulty, of course, is that I cannot have Jones's feeling at time t_1—unless I *am* Jones, and the time *is* t_1.

From this simple observation, certain things follow. For example, the account according to which the *intension* of the word 'pain' is a certain *quality* which 'I know from my own case' must be wrong. But this is not to refute dualism, since the dualist need not maintain that I know the intension of the English word 'pain' from my own case, but only that I experience the referent of the word.

What then is the intension of 'pain'? I am inclined to say that 'pain' is a cluster-concept. That is, the application of the word 'pain' is controlled by a whole cluster of criteria, *all of which can be regarded as synthetic.*[3] As a consequence, there is no satisfactory way of answering the question 'What does "pain" mean?' except by giving an exact synonym (e.g. 'Schmerz'); but there are a million and one different ways of saying what pain *is*. One can, for example, say that pain is that feeling which is normally evinced by saying 'ouch', or by wincing, or in a variety of other ways (or often not evinced at all).

All this is compatible with logical behaviorism. The logical behaviorist would reply: 'Exactly. "Pain" is a cluster-concept —that is to say, it stands for *a cluster of phenomena.*' But that is not what I mean. Let us look at another kind of cluster-concept (cluster-concepts, of course, are not a homogeneous class): names of diseases.

We observe that, when a virus origin was discovered for polio, doctors said that certain cases in which all the symptoms of polio had been present, but in which the virus had been absent, had turned out not to be cases of polio at all. Similarly, if a virus should be discovered which normally (almost invariably) is the cause of what we presently call 'multiple sclerosis', the hypothesis that this virus is *the* cause of multiple sclerosis would not be falsified if,

in some few exceptional circumstances, it was possible to have all the symptoms of multiple sclerosis for some other combination of reasons, or if this virus caused symptoms not presently recognized as symptoms of multiple sclerosis in some cases. These facts would certainly lead the lexicographer to *reject* the view that 'multiple sclerosis' means 'the simultaneous presence of such and such symptoms'. Rather he would say that 'multiple sclerosis' means 'that disease which is normally responsible for some or all of the following symptoms . . .'

Of course, he does not have to say this. Some philosophers would prefer to say that 'polio' *used to mean* 'the simultaneous presence of such-and-such symptoms'. And they would say that the *decision* to accept the presence or absence of a virus as a criterion for the presence or absence of polio represented a *change of meaning*. But this runs strongly counter to our common sense. For example, doctors used to say 'I believe polio is caused by a virus'. On the 'change of meaning' account, those doctors were *wrong*, not *right*. Polio, *as the word was then used*, was not always caused by a virus; it is only what *we* call polio that is always caused by a virus. And if a doctor ever said (and many did) 'I believe this may not be a case of polio', knowing that all of the textbook symptoms were present, that doctor must have been contradicting himself (even if we, to-day, would say that he was right) or, perhaps, 'making a disguised linguistic proposal'. Also, this account runs counter to good linguistic methodology. The definition we proposed a paragraph back—'multiple sclerosis' means 'the disease that is normally *responsible* for the following symptoms . . .' —has an exact analogue in the case of polio. This kind of definition leaves open the question whether there is a single cause or several. It is consonant with such a definition to speak of 'discovering a single origin for polio (or two or three or four)', to speak of 'discovering *X* did not have polio' (although he exhibited all the symptoms of polio), and to speak of 'discovering *X* did have polio' (although he exhibited *none* of the 'textbook symptoms'). And, finally, such a definition does not require us to say that any 'change of meaning' took place. Thus, this is surely the definition that a good lexicographer would adopt. But this entails *rejecting* the 'change of meaning' account as a philosopher's invention.[4]

Accepting that this is the correct account of the names of diseases, what follows? There *may* be analytic entailments connecting diseases and symptoms (although I shall argue against this). For example, it looks plausible to say that:

'Normally people who have multiple sclerosis have some or all of the following symptoms . . .' is a necessary ('analytic') truth. But it does not follow that 'disease talk' is translatable into 'symptom talk'. Rather the contrary follows (as is already indicated by the presence of the word 'normally'): statements about multiple sclerosis are not translatable into statements about the symptoms of multiple sclerosis, not because disease talk is 'systematically ambiguous' and symptom talk is 'specific', but because *causes* are not logical constructions out of their *effects*.

In analogy with the foregoing, both the dualist and the materialist would want to argue that, although the meaning of 'pain' may be *explained* by reference to overt behavior, what we mean by 'pain' is not the presence of a cluster of responses, but rather the presence of an event or condition that normally causes those responses. (Of course the pain is not the whole cause of the pain behavior, but only a suitably invariant part of that cause,[5] but, similarly, the virus-caused tissue damage is not the whole cause of the individual symptoms of polio in some individual case, but a suitably invariant part of the cause.) And they would want to argue further, that even if it *were* a nec-

essary truth that

'Normally, when one says "ouch" one has a pain'

or a necessary truth that

'Normally, when one has a pain one says "ouch" '

this would be an interesting observation about what 'pain' means, but it would shed no metaphysical light on what pain *is* (or *isn't*). And it certainly would not follow that 'pain talk' is translatable into 'response talk', or that the failure of translatability is only a matter of the 'systematic ambiguity' of pain talk as opposed to the 'specificity' of response talk: quite the contrary. Just as before, *causes* (pains) are *not* logical constructions out of their *effects* (behavior).

The traditional dualist would, however, want to go farther, and deny the *necessity* of the two propositions just listed. Moreover, the traditional dualist is right: there is nothing self-contradictory, as we shall see below, in talking of hypothetical worlds in which there are pains but *no* pain behavior.

The analogy with names of diseases is still preserved at this point. Suppose I identify multiple sclerosis as the disease that normally produces certain symptoms. If it later turns out that a certain virus is the cause of multiple sclerosis, using this newly discovered criterion I may then go on to find out that multiple sclerosis has quite different symptoms when, say, the average temperature is lower. I can then perfectly well talk of a hypothetical world (with lower temperature levels) in which multiple sclerosis does *not* normally produce the usual symptoms. It is true that if the *words* 'multiple sclerosis' are used in any world in such a way that the above lexical definition is a good one, *then* many victims of the disease must have had some or all of the following symptoms . . . And in the same way it is true that *if* the explanation suggested of the word 'pain' is a good one (i.e. 'pain is the feeling that is normally being evinced when someone says "ouch", or winces, or screams, etc.'), *then* persons in pain must have at some time winced or screamed or said 'ouch'—but this does *not* imply that 'if someone ever had a pain, then someone must at some time have winced or screamed or said "ouch".' To conclude this would be to confuse preconditions for *talking* about pain as *we* talk about pain with preconditions for the existence of pain.

The analogy we have been developing is not an identity: linguistically speaking, mind words and names of diseases are different in a great many respects. In particular, *first person uses* are very different: a man may have a severe case of polio and not know it, even if he knows the word 'polio', but one cannot have a severe pain and not know it. At first blush, this may look like a point in favor of logical behaviorism. The logical behaviorist may say: it is because the premisses 'John says he has a pain', 'John knows English', and 'John is speaking in all sincerity',[6] *entail* 'John has a pain', that pain reports have this sort of special status. But even if this is right, it does not follow that logical behaviorism is correct unless *sincerity* is a 'logical construction out of overt behavior'! A far more reasonable account is this: one can have a 'pink elephant hallucination', but one cannot have a 'pain hallucination', or an 'absence of pain hallucination', simply because any situation that a person cannot discriminate from a situation in which he himself has a pain *counts* as a situation in which he has a pain, whereas a situation that a person cannot distinguish from one in which a pink elephant is present does not necessarily *count* as the presence of a pink elephant.

To sum up: I believe that pains are not clusters of responses, but that they are (normally, in our experience to date) the causes of certain clusters of responses.

Moreover, although this is an empirical fact, it underlies the possibility of talking about pains in the particular way in which we do. However, it does not rule out in any way the possibility of worlds in which (owing to a difference in the environmental and hereditary conditions) pains are not responsible for the usual responses, or even are not responsible for any responses at all.

Let us now engage in a little science fiction. Let us try to describe some worlds in which pains are related to responses (and also to causes) in quite a different way than they are in our world.

If we confine our attention to nonverbal responses by full grown persons, for a start, then matters are easy. Imagine a community of 'super-spartans' or 'super-stoics'—a community in which the adults have the ability to successfully suppress *all* involuntary pain behavior. They may, on occasion, admit that they feel pain, but always in pleasant well-modulated voices —even if they are undergoing the agonies of the damned. They do *not* wince, scream, flinch, sob, grit their teeth, clench their fists, exhibit beads of sweat, or otherwise act like people in pain or people suppressing the unconditioned responses associated with pain. However, they do feel pain, and they dislike it (just as we do). They even admit that it takes a great effort of will to behave as they do. It is only that they have what they regard as important ideological reasons for behaving as they do, and they have, through years of training, learned to live up to their own exacting standards.

It may be contended that children and not fully mature members of this community will exhibit, to varying degrees, normal unconditioned pain behavior, and that this is all that is necessary for the ascription of pain. On this view, the *sine qua non* for significant ascription of pain to a species is that its immature members should exhibit unconditioned pain responses.

One might well stop to ask whether this statement has even a clear meaning. Supposing that there are Martians: do we have any criterion for something being an 'unconditioned pain response' for a Martian? Other things being equal, one *avoids* things with which one has had painful experiences: this would suggest that *avoidance* behavior might be looked for as a universal unconditioned pain response. However, even if this were true, it would hardly be specific enough, since avoidance can also be an unconditioned response to many things that we do not associate with pain—to things that disgust us, or frighten us, or even merely bore us.

Let us put these difficulties aside, and see if we can devise an imaginary world in which there are not, even by lenient standards, any unconditioned pain responses. Specifically, let us take our 'super-spartans', and let us suppose that after millions of years they begin to have children who are born fully acculturated. They are born speaking the adult language, knowing the multiplication table, having opinions on political issues, and *inter alia* sharing the dominant spartan beliefs about the importance of not evincing pain (except by way of verbal report, and even that in a tone of voice that suggests indifference). Then there would not *be* any 'unconditioned pain responses' in this community (although there might be unconditioned *desires* to make certain responses—desires which were, however, always suppressed by an effort of will). Yet there is a clear absurdity to the position that one cannot ascribe to these people a capacity for feeling pain.

To make this absurdity evident, let us imagine that we succeed in converting an adult 'super-spartan' to *our* ideology. Let us suppose that he begins to evince pain in the normal way. Yet he reports that the pains he is feeling are not more *intense* than are the ones he experienced prior to conversion—indeed, he may say that giving expression to them makes

them *less* intense. In this case, the logical behaviorist would have to say that, through the medium of this one member, we had demonstrated the existence of unconditioned pain responses in the whole species, and hence that ascription of pain to the species is 'logically proper'. But this is to say that had this one man never lived, and had it been possible to demonstrate only indirectly (via the use of *theories*) that these beings feel pain, then pain ascriptions *would* have been improper.

We have so far been constructing worlds in which the relation of pain to its nonverbal *effects* is altered. What about the relation of pain to *causes?* This is even more easy for the imagination to modify. Can one not imagine a species who feel pain only when a magnetic field is present (although the magnetic field causes no detectable damage to their bodies or nervous systems)? If we now let the members of such a species become converts *to* 'super-spartanism', we can depict to ourselves a world in which pains, in our sense, are clearly present, but in which they have neither the normal causes nor the normal effects (apart from verbal reports).

What about verbal reports? Some behaviorists have taken these as the characteristic form of pain behavior. Of course, there is a difficulty here: If 'I am in pain' means 'I am disposed to utter this kind of verbal report' (to put matters crudely), then how do we tell that any particular report is 'this kind of verbal report'? The usual answer is in terms of the unconditioned pain responses and their assumed supplantation by the verbal reports in question. However, we have seen that there are no *logical* reasons for the existence of unconditioned pain responses in all species capable of feeling pain (there *may* be logical reasons for the existence of avoidance desires, but avoidance *desires* are not themselves behavior any more than pains are).

Once again, let us be charitable to the extent of waiving the first difficulty that comes to mind, and let us undertake the task of trying to imagine a world in which there are not even pain *reports*. I will call this world the '*X*-world'. In the *X*-world we have to deal with 'super-super-spartans'. These have been super-spartans for so long, that they have begun to suppress even *talk* of pain. Of course, each individual *X*-worlder may have his private way of thinking about pain. He may even have the *word* 'pain' (as before, I assume that these beings are born fully acculturated). He may *think* to himself: 'This pain is intolerable. If it goes on one minute longer I shall scream. Oh No! I mustn't do that! That would disgrace my whole family . . .' But *X*-worlders do not even admit to *having* pains. They pretend not to know either the word or the phenomenon to which it refers. In short, if pains are 'logical constructs out of behavior', then our *X*-worlders behave so as not to have pains!—Only, of course, they do have pains, and they know perfectly well that they have pains.

If this last fantasy is not, in some disguised way, self-contradictory, then logical behaviorism is simply a mistake. Not only is the second thesis of logical behaviorism—the existence of a near-translation of pain talk into behavior talk—false, but so is even the first thesis—the existence of 'analytic entailments'. Pains *are* responsible for certain kinds of behavior—but only in the context of our beliefs, desires, ideological attitudes, and so forth. From the statement '*X* has a pain' by itself *no* behavioral statement follows—not even a behavioral statement with a 'normally' or a 'probably' in it.

In our concluding section we shall consider the logical behaviorist's stock of counter-moves to this sort of argument. If the logical behaviorist's positive views are inadequate owing to an oversimplified view of the nature of cluster words—amounting, in some instances, to an open denial that it is *possible* to have a word

governed by a cluster of indicators, *all* of which are synthetic—his negative views are inadequate owing to an oversimplified view of empirical reasoning. It is unfortunately characteristic of modern philosophy that its problems should overlap three different areas—to speak roughly, the areas of linguistics, logic, and 'theory of theories' (scientific methodology)—and that many of its practitioners should try to get by with an inadequate knowledge of at least two out of the three.

Some Behaviorist Arguments

We have been talking of 'X-worlders' and 'super-spartans'. No one denies that, in *some* sense of the term, such fantasies are 'intelligible'. But 'intelligibility' can be a superficial thing. A fantasy may be 'intelligible', at least at the level of 'surface grammar', although we may come to see, on thinking about it for a while, that some absurdity is involved. Consider, for example, the supposition that last night, just on the stroke of midnight, all distances were instantaneously doubled. Of course, we did not notice the change, for *we* ourselves also doubled in size! This story may seem intelligible to us at first blush, at least as an amusing possibility. On reflection, however, we come to see that logical contradiction is involved. For 'length' means nothing more nor less than a relation to a standard, and it is a contradiction to maintain that the length of everything doubled, while the relations to the standards remained unchanged.

What I have just said (speaking as a logical behaviorist might speak) is false, but not totally so. It is false (or at least the last part is false), because 'length' does *not* mean 'relation to a standard'. If it did (assuming a 'standard' has to be a macroscopic material object, or anyway a material object), it would make no sense to speak of distances in a world in which there were only gravitational and electromagnetic fields, but no material objects. Also, it would make no sense to speak of the *standard* (whatever it might be) as having changed its length. Consequences so counter-intuitive have led many physicists (and even a few philosophers of physics) to view 'length' not as something operationally defined, but as a theoretical magnitude (like electrical charge), which can be measured in a virtual infinity of ways, but which is not explicitly and exactly definable in terms of any of the ways of measuring it. Some of these physicists —the 'unified field' theorists—would even say that, far from it being the case that 'length' (and hence 'space') depends on the existence of suitably related material bodies, material bodies are best viewed as local variations in the curvature of space —that is to say, local variations in the intensity of a certain magnitude (the tensor g_{ik}), one aspect of which we experience as 'length'.

Again, it is far from true that the hypothesis 'last night, on the stroke of midnight, everything doubled in length' has no testable consequences. For example, if last night everything did double in length, and the velocity of light did not also double, then this morning we would have experienced an apparent halving of the speed of light. Moreover, if g (the gravitational constant) did not double, then we would have experienced an apparent halving in the intensity of the gravitational field. And if h (Planck's constant) did not change, then . . . In short, our world would have been bewilderingly different. And if we could survive at all, under so drastically altered conditions, no doubt some clever physicist would figure out what had happened.

I have gone into such detail just to make the point that in philosophy things are rarely so simple as they seem. The 'doubling universe' is a favorite classroom example of a 'pseudo-hypothesis'—yet it is the worst possible example if a 'clear case' is desired. In the first place, what is desired is a hypothesis with no testable consequences—yet *this* hypothesis, as it

is always stated, *does* have testable consequences (perhaps some more complex hypothesis does not; but then we have to see this more complex hypothesis stated before we can be expected to discuss it). In the second place, the usual argument for the absurdity of this hypothesis rests on a simplistic theory of the meaning of 'length' —and a full discussion of *that* situation is hardly possible without bringing in considerations from unified field theory and quantum mechanics (the latter comes in connection with the notion of a 'material standard'). But, the example aside, one can hardly challenge the point that a superficially coherent story may contain a hidden absurdity.

Or can one? Of course, a superficially coherent story may contain a hidden contradiction, but the whole point of the logical behaviorist's sneering reference to 'surface grammar' is that *linguistic coherence, meaningfulness of the individual terms,* and *logical consistency,* do not by themselves guarantee freedom from another kind of absurdity—there are 'depth absurdities' which can only be detected by more powerful techniques. It is fair to say that to-day, after thirty years of this sort of talk, we lack both a single *convincing* example of such a depth absurdity, and a technique of detection (or alleged technique of detection) which does not reduce to 'untestable, *therefore* nonsense'.

To come to the case at hand: the logical behaviorist is likely to say that our hypothesis about 'X-worlders' is untestable in principle (if there *were* 'X-worlders', by hypothesis we couldn't distinguish them from people who really didn't know what pain is); and *therefore* meaningless (apart from a certain 'surface significance' which is of no real interest). If the logical behaviorist has learned a little from 'ordinary language philosophy', he is likely to shy away from saying 'untestable, therefore *meaningless*', but he is still likely to say or at least think: 'untestable, therefore in *some* sense absurd'. I shall try to meet this 'argument' *not* by challenging the premiss, be it overt or covert, that 'untestable synthetic statement' is some kind of contradiction in terms (although I believe that premiss to be mistaken), but simply by showing that, on any but the most naive view of testability, our hypothesis *is* testable.

Of course, I could not do this if it were true that 'by hypothesis, we couldn't distinguish X-worlders from people who *really* didn't know what pain is'. But that isn't true—at any rate, it isn't true 'by hypothesis'. What is true by hypothesis is that we couldn't distinguish X-worlders from people who really didn't know what pain is *on the basis of overt behavior alone.* But that still leaves many other ways in which we might determine what is going on 'inside' the X-worlders—in both the figurative and literal sense of 'inside'. For example, we might examine their *brains.*

It is a fact that when pain impulses are 'received' in the brain, suitable electrical detecting instruments record a characteristic 'spike' pattern. Let us express this briefly (and too simply) by saying that 'brain spikes' are one-to-one correlated with experiences of pain. If our X-worlders belong to the human species, then we can verify that they do feel pains, notwithstanding their claim that they don't have any idea what pain is, by applying our electrical instruments and detecting the tell-tale 'brain spikes'.

This reply to the logical behaviorist is far too simple to be convincing. 'It is true,' the logical behaviorist will object, 'that experiences of pain are one-to-one correlated with "brain spikes" in the case of normal human beings. But you don't know that the X-worlders are normal human beings, in this sense—in fact, you have every reason to suppose that they are *not* normal human beings'. This reply shows that no *mere* correlation, however carefully verified in the case of normal human beings, can be used to verify

ascriptions of pain to X-worlders. Fortunately, we do not have to suppose that our knowledge will always be restricted to mere correlations, like the pain-'brain spike' correlation. At a more advanced level, considerations of simplicity and coherence can begin to play a role in a way in which they cannot when only crude observational regularities are available.

Let us suppose that we begin to detect waves of a new kind, emanating from human brains—call them 'V-waves'. Let us suppose we develop a way of 'decoding' V-waves so as to reveal people's unspoken thoughts. And, finally, let us suppose that our 'decoding' technique also works in the case of the V-waves emanating from the brains of X-worlders. How does this correlation differ from the pain-'brain spike' correlation?

Simply in this way: it is reasonable to say that 'spikes'—momentary peaks in the electrical intensity in certain parts of the brain—could have almost any cause. But waves which go over into coherent English (or any other language), under a relatively simple decoding scheme, could not have just any cause. The 'null hypothesis'—that this is just the operation of 'chance'—can be dismissed at once. And if, in the case of human beings, we verify that the decoded waves correspond to what we are in fact thinking, then the hypothesis that this same correlation holds in the case of X-worlders will be assigned an immensely high probability, simply because no other likely explanation readily suggests itself. But 'no other likely explanation readily suggests itself' isn't verification, the logical behaviorist may say. On the contrary. How, for example, have we verified that cadmium lines in the spectrographic analysis of sunlight indicate the presence of cadmium in the sun? Mimicking the logical behaviorist, we might say: 'We have verified that under normal circumstances, cadmium lines only occur when heated cadmium is present. But we don't know that circumstances on the sun are normal in this sense'. If we took this seriously, we would have to *heat cadmium on the sun* before we could say that the regularity upon which we base our spectrographic analysis of sunlight had been verified. In fact, we have verified the regularity under 'normal' circumstances, and we can *show* (deductively) that *if* many other laws, that have also been verified under 'normal' circumstances and *only* under 'normal' circumstances (i.e. never on the surface of the sun), hold on the sun, *then* this regularity holds also under 'abnormal' circumstances. And if someone says, 'But perhaps *none* of the usual laws of physics hold on the sun', we reply that this is like supposing that a random process always produces coherent English. The fact is that the 'signals' (sunlight, radio waves, etc.) which we receive from the sun cohere with a vast body of theory. Perhaps there is some other explanation than that the sun obeys the usual laws of physics; but *no other likely explanation suggests itself.* This sort of reasoning *is* scientific verification; and if it is not reducible to simple Baconian induction—well, then, philosophers must learn to widen their notions of verification to embrace it.

The logical behaviorist might try to account for the decodability of the X-worlders' 'V-waves' into coherent English (or the appropriate natural language) without invoking the absurd 'null hypothesis'. He might suggest, for example, that the 'X-worlders' are having fun at our expense—they are able, say, to produce misleading V-waves at will. If the X-worlders have brains quite unlike ours, this may even have some plausibility. But once again, in an advanced state of knowledge, considerations of coherence and simplicity may quite conceivably 'verify' that this is false. For example, the X-worlders may have brains quite like ours, rather than unlike ours. And we may have built up enough theory to say how the brain of a human being should 'look' if that human being were pretending not to

be in pain when he was, in fact, in pain. Now consider what the 'misleading *V*-waves' story requires: it requires that the *X*-worlders produce *V*-waves in quite a different way than we do, without specifying what that different way is. Moreover, it requires that this be the case, although the reverse hypothesis—that *X*-worlders' brains function *exactly* as human brains do—in fact, that they *are* human brains—fits all the data. Clearly, this story is in serious methodological difficulties, and any other 'counter-explanation' that the logical behaviorist tries to invoke will be in similar difficulties. In short, the logical behaviorist's argument reduces to this: 'You cannot verify "psycho-physical" correlations in the case of *X*-worlders (or at least, you can't verify ones having to do, directly or indirectly, with *pain*), because, by hypothesis, *X*-worlders won't tell you (or indicate behaviorally) when they are in pain'. 'Indirect verification'—verification using theories which have been 'tested' only in the case of human beings—is not verification at all, because *X*-worlders *may* obey different laws than human beings. And it is not incumbent upon *me* (the logical behaviorist says) to suggest what those laws might be: it is incumbent upon *you* to rule out *all* other explanations. And this is a silly argument. The scientist does not have to rule out all the ridiculous theories that someone *might* suggest; he only has to show that he has ruled out any reasonable alternative theories that one might put forward on the basis of present knowledge.

Granting, then, that we might discover a technique for 'reading' the unspoken thoughts of *X*-worlders: we would then be in the same position with respect to the *X*-worlders as we were with respect to the original 'super-spartans'. The super-spartans were quite willing to tell us (and each other) about their pains; and we could see that their pain talk was linguistically coherent and situationally appropriate (e.g. a super-spartan will tell you that he feels intense pain when you touch him with a red hot poker). On this basis, we were quite willing to grant that the super-spartans did, indeed, feel pain—all the more readily, since the deviancy in their behavior had a perfectly convincing ideological explanation. (Note again the role played here by considerations of coherence and simplicity.) But the *X*-worlders also 'tell' us (and, perhaps, each other), exactly the same things, albeit *un*willingly (by the medium of the involuntarily produced '*V*-waves'). Thus we have to say—at least, we have to say as long as the '*V*-wave' theory has not broken down—that the *X*-worlders are what they, in fact, are—just 'super-super-spartans'.

Let us now consider a quite different argument that a logical behaviorist might use. 'You are assuming,' he might say, 'the following principle:

> If someone's brain is in the same state as that of a human being in pain (not just at the moment of the pain, but before and after for a sufficient interval), then he is in pain. Moreover, this principle is one which it would never be reasonable to give up (on your conception of 'methodology'). Thus, you have turned it into a tautology. But observe what turning this principle into a tautology involves: it involves changing the meaning of 'pain'. What 'pain' means for *you* is: the presence of pain, in the colloquial sense of the term, *or* the presence of a brain state identical with the brain state of someone who feels pain. Of course, in that sense we can verify that your 'X-worlders' experience 'pain'—but that is not the sense of 'pain' at issue.

The reply to this argument is that the premiss is simply false. It is just not true that, on my conception of verification, it would *never* be reasonable to give up the principle stated. To show this, I have to beg your pardons for engaging in a little more science fiction. Let us suppose that scientists discover yet another kind of waves—call them '*W*-waves'. Let us suppose that *W*-waves do not emanate from human brains, but that they are detected

emanating from the brains of *X*-worlders. And let us suppose that, once again, there exists a simple scheme for decoding *W*-waves into coherent English (or whatever language *X*-worlders speak), and that the 'decoded' waves 'read' like this: 'Ho, ho! are we fooling those Earthians! They think that the *V*-waves they detect represent our thoughts! If they only knew that instead of pretending not to have pains when we really have pains, we are really pretending to pretend not to have pains when we really do have pains when we really don't have pains!' Under these circumstances, we would 'doubt' (to put it mildly) that the same psycho-physical correlations held for normal humans and for *X*-worlders. Further investigations might lead us to quite a number of different hypotheses. For example, we might decide that *X*-worlders don't think with their brains at all—that the 'organ' of thought is not just the brain, in the case of *X*-worlders, but some larger structure—perhaps even a structure which is not 'physical' in the sense of consisting of elementary particles. The point is that what is necessarily true is not the principle stated in the last paragraph, but rather the principle:

> If someone (some organism) is in the same state as a human being in pain in all relevant respects, then he (that organism) is in pain.

—And *this* principle *is* a tautology by anybody's lights! The only *a priori* methodological restriction I am imposing here is this one:

> If some organism is in the same state as a human being in pain in all respects *known* to be relevant, and there is no reason to suppose that there exist *un*known relevant respects, then don't postulate any.

—But this principle is not a 'tautology'; in fact, it is not a *statement* at *all*, but a methodological directive. And deciding to conform to this directive is not (as hardly needs to be said) changing the meaning of the word 'pain', or of *any* word.

There are two things that the logical behaviorist can do: he can claim that ascribing pains to *X*-worlders, or even super-spartans, involves a 'change of meaning',[7] or he can claim that ascribing pains to super-spartans, or at least to *X*-worlders, is 'untestable'. The first thing is a piece of unreasonable linguistics; the second, a piece of unreasonable scientific method. The two are, not surprisingly, mutually supporting: the unreasonable scientific method makes the unreasonable linguistics appear more reasonable. Similarly, the normal ways of thinking and talking are mutually supporting: reasonable linguistic field techniques are, needless to say, in agreement with reasonable conceptions of scientific method. Madmen sometimes have consistent delusional systems; so madness and sanity can both have a 'circular' aspect. I may not have succeeded, in this paper, in breaking the 'delusional system' of a committed logical behaviorist; but I hope to have convinced the uncommitted that that system need not be taken seriously. If we have to choose between 'circles', the circle of reason is to be preferred to any of the many circles of unreason.

Notes

1. E.g. these two points are fairly explicitly stated in Strawson's *Individuals*. Strawson has told me that he no longer subscribes to point (1), however.

2. An attempted fourth alternative—i.e. an alternative to dualism, materialism, *and* behaviorism—is sketched in "The Mental Life of Some Machines," which appeared in the Proceedings of the Wayne Symposium on the Philosophy of Mind. This fourth alternative is materialistic in the wide sense of being compatible with the view that organisms, including human beings, are physical systems consisting of elementary particles and obeying the laws of physics, but does not require that such 'states' as *pain* and *preference* be defined in a way which makes reference to either overt behavior or physical-chemical constitution. The idea, briefly, is that predicates which apply to

a system by virtue of its *functional organization* have just this characteristic: a given functional organization (e.g. a given inductive logic, a given rational preference function) may realize itself in almost any kind of overt behavior, depending upon the circumstances, and is capable of being 'built into' structures of many different logically possible physical (or even metaphysical) constitutions. Thus the statement that a creature prefers *A* to *B* does not tell us whether the creature has a carbon chemistry, or a silicon chemistry, or is even a disembodied mind, nor does it tell us how the creature would behave under any circumstances specifiable without reference to the creature's other preferences and beliefs, but it does not thereby become something 'mysterious'.

3. I mean not only that *each* criterion can be regarded as synthetic, but also that the cluster is *collectively* synthetic, in the sense that we are free in certain cases to say (for reason of inductive simplicity and theoretical economy) that the term applies although the whole cluster is missing. This is completely compatible with saying that the cluster serves to fix the meaning of the word. The point is that when we specify something by a cluster of indicators we assume that people will *use their brains*. That criteria may be over-ridden when good sense demands is the sort of thing we may regard as a 'convention associated with discourse' (Grice) rather than as something to be stipulated in connection with the individual words.

4. Cf. "Dreaming and 'Depth Grammar,' " *Analytical Philosophy*, vol. 1.

5. Of course, 'the cause' is a highly ambiguous phrase. Even if it is correct in certain contexts to say that certain events in the brain are 'the cause' of my pain behavior, it does *not* follow (as has sometimes been suggested) that my pain must be 'identical' with these neural events.

6. This is suggested in Wittgenstein's *Philosophical Investigations*.

7. This popular philosophical move is discussed in "Dreaming and 'Depth Grammar,' " *Analytical Philosophy*, vol. 1.