SYDNEY SHOEMAKER

FUNCTIONALISM AND QUALIA

(Received 11 June, 1974)

1. In their recent paper 'What Psychological States are Not' N. J. Block and J. A. Fodor raise a number of objections to the 'functional state identity theory' (FSIT), which says that "for any organism that satisfies psychological predicates at all, there exists a unique best *description* such that each psychological state of the organism is identical with one of its machine states relative to that description."[1] FSIT is a version of 'functionalism', which they characterize as the more general doctrine that "the type-identity conditions for psychological states refer only to their relations to inputs, outputs, and one another."[2] Most of the objections Block and Fodor raise they take to be objections only to FSIT, and not to functionalism more broadly construed. I shall not be concerned with these objections here. But they raise one objection which, they say, "might be taken to show that psychological states cannot be functionally defined *at all* and that they cannot be put into correspondence with *any* properties definable over abstract automata."[3] Briefly put, the objection is that the way of 'type-identifying' psychological states proposed by FSIT, and by functionalism generally, "fails to accommodate a feature of at least some such states that is critical for determining their type: namely their 'qualitative' character."[4]

Block and Fodor devote only a couple of pages to this objection, and raise it in a fairly tentative way; so it is quite likely that the length of my discussion of it here is disproportionate to the importance they put on it. But they have given a concise and vivid formulation to an objection which is felt, and voiced in conversation, more often than it is expressed in print, and which seems to me to raise fundamental issues. Other philosophers have raised much the same objection by saying that functionalism (or behaviorism, or materialism, or 'causal' theories of the mind – the objection has been made against all of these) cannot account for the 'raw feel' component of mental states, or for their 'internal', or

Philosophical Studies **27** (1975) 291–315.

'phenomenological', character. My primary concern here is not with whether this objection is fatal to FSIT; if I understand that theory correctly, it is sufficiently refuted by the other objections Block and Fodor raise against it. But as they characterize functionalism 'in the broad sense', it is, while vague, a view which many philosophers, myself included, find attractive; and it seems to me worth considering whether it can be defended against this objection.

I shall follow Block and Fodor in speaking of mental states (or rather, of some mental states) as having 'qualitative character(s)' or 'qualitative content'. I hope that it will emerge in the ensuing discussion that this does not commit me to anything which a clear headed opponent of 'private objects', or of 'private language', should find objectionable.

2. Block and Fodor develop their objection in two stages. The first of these they call the 'inverted qualia argument', and the second can be called the 'absent qualia argument'.

Because they are unpersuaded by the familiar 'verificationist' arguments against the conceptual coherence of the 'inverted spectrum hypothesis', Block and Fodor are inclined to think that cases of 'inverted qualia' may be possible. They take it that there would be qualia inversion (presumably an extreme case of it) if it were true that "every person does, in fact, have slightly different qualia (or, better still, grossly different qualia) when in whatever machine table state is alleged to be identical to pain."[5] The possibility of this is incompatible with functionalism on the plausible assumption that "nothing would be a token of the type 'pain state' unless it felt like a pain, ... even if it were connected to all of the other psychological states of the organism in whatever ways pains are."[6]

Block and Fodor do not regard the possibility of qualia inversion as constituting by itself a decisive objection to functionalism, for they think that it may be open to the functionalist to deny the *prima facie* plausible assumption that pains must be qualitatively similar (and, presumably, the related assumption that anything qualitatively identical to a pain is itself a pain).[7] If qualia inversion actually occurred in the case of pain (i.e., if a state functionally identical to a pain differed from it in qualitative character), then, they say, "it might be reasonable to say that the character of an organism's qualia is irrelevant to whether it is in pain or

(equivalently) that pains feel quite different to different organisms."[8] Such a view is not in fact unheard of. According to Don Locke, "A sensation's being a pain sensation is not a matter of how it feels, but a matter of its being of the sort caused by bodily damage and leading to pain behavior."[9] And Alan Donagan has attributed to Wittgenstein the view that "you and I correctly say that we have the same sensation, say toothache, if we both have something frightful that we would naturally express by holding and rubbing our jaws, by certain kinds of grimace, and the like. Whether the internal character of what is expressed in these ways is the same for you as for me is irrelevant to the meaning of the word 'toothache'."[10]

But while Block and Fodor do not dismiss this response to the inverted qualia argument as obviously mistaken, they see it as possibly opening the door to an argument much more damaging to functionalism, namely the absent qualia argument. Their thought may be that once it is admitted that a given functional state can exist without having a given 'qualitative content', it will be difficult to deny the possibility that it might exist without having any qualitative content (or character) at all. At any rate, they go on to say that

> For all that we know, it may be nomologically possible for two psychological states to be functionally identical (that is, to be identically connected with inputs, outputs, and successor states), even if only one of the states has a qualitative content. In this case, FSIT would require us to say that an organism might be in pain even though it is feeling *nothing at all*, and this consequence seems totally unacceptable.[11]

And if cases of 'absent qualia' are possible, i.e., if a state can be functionally identical to a state having a qualitative character without itself having a qualitative content, then not only FSIT, but also functionalism in the broad sense, would seem to be untenable.

3. If mental states can be alike or different in 'qualitative character', we should be able to speak of a class of states, call them 'qualitative states', whose 'type-identity conditions' could be specified in terms of the notion of qualitative (or 'phenomenological') similarity. For each determinate qualitative character a state can have, there is (i.e., we can define) a determinate qualitative state which a person has just in case he has a state having precisely that qualitative character. For example, there is a qualitative state someone has just in case he has a sensation

that feels exactly the way my most recent headache felt. Qualitative states will presumably be 'mental', or 'psychological', states. And this calls into question the suggestion of Block and Fodor that a functionalist could deal with the 'inverted qualia argument' by maintaining that 'the character of an organism's qualia is irrelevant to whether it is in pain'. It would of course be self-contradictory to hold that the character of an organism's qualia is irrelevant to what qualitative states it has. And Block and Fodor are presumably committed to the claim that qualitative states cannot themselves be functionally defined, or at least that this is so if there can be cases of 'inverted qualia'. For if qualitative states could be functionally defined, the fact that mental states have qualitative character would provide no problem for functionalism. Thus (assuming the possibility of qualia inversion) there will be one class of mental states, namely the qualitative states themselves, that cannot be functionally defined.

This raises questions which I shall return to in later sections, namely (a) in what sense are qualitative states not functionally definable (or, in what sense are they not functionally definable if qualia inversion is possible), and (b) is their being functionally undefinable (in whatever sense they are) seriously damaging to functionalism? As we shall see in the remainder of the present section, this question is also raised by a consideration of the alleged possibility of 'absent qualia'.

We can establish the impossibility of cases of 'absent qualia' if we can show that if a state is functionally identical to a state having qualitative content then it must itself have qualitative content. One might try to do this by construing the notion of functional identity in such a way that qualitative states are included among the 'other psychological states' by relation to which, along with input and output, the 'type-identity' of a given psychological state is to be defined. Thus one might argue that if a given psychological state has a certain qualitative character, this involves its standing in some determinate relationship to some particular qualitative state (namely the qualitative state a person is in just in case he is in a state having that qualitative character), and that any state functionally identical to it must stand in the same relationship to that qualitative state, and so must have the same qualitative character.[12] But this argument is not very convincing. One objection that is likely to be made to it is that since qualitative states cannot themselves be function-

ally defined (assuming the possibility of *inverted* qualia), it is illegitimate to include them among the psychological states by reference to which other psychological states are functionally defined, or in terms of which 'functional identity' is defined. I shall return to this objection later, since it is also a *prima facie* objection against the more plausible argument I shall present next. Another objection is that the relationship which a state has to a qualitative state, in having the 'qualitative character' corresponding to that qualitative state, is not anything like a causal relationship and so is not the sort of relationship in terms of which a psychological state can be functionally defined. But the argument I shall present next is not open to this objection, and does seem to me to show that on any plausible construal of the notion of functional identity a state cannot be functionally identical to a state having qualitative character without itself having qualitative character.

One important way in which pains are related to other psychological states is that they give rise, under appropriate circumstances, to introspective awareness of themselves as having certain qualitative characters, i.e., as feeling certain ways. I shall assume that the meaning of this can be partially unpacked by saying that being in pain typically gives rise, given appropriate circumstances, to what I shall call a 'qualitative belief', i.e., a belief to the effect that one feels a certain way (or, more abstractly, is in a state having a certain qualitative character, or, in still other terms, has a certain qualitative state). Any state functionally identical to a pain state will share with the pain state not only (1) its tendency to influence overt behavior in certain ways, and (2) its tendency to produce in the person the belief that there is something organically wrong with him (e.g., that he has been cut or burnt), but also (3) its tendency to produce qualitative beliefs in the person, i.e., to make him think that he has a pain having a certain qualitative character (one that he dislikes). According to the 'absent qualia argument', such a state may nevertheless lack qualitative character, and so fail to be a pain. Let us consider whether this is plausible.

Supposing such cases of 'absent qualia' are possible, how might we detect such a case if it occurred? And with what right does each of us reject the suggestion that perhaps his own case is such a case, and that he himself is devoid of states having qualitative character? Indeed, with what right do we reject the suggestion that perhaps no one ever has any

feelings (or other states having qualitative character) at all? It is, of course, a familiar idea that behavior provides inconclusive evidence as to what qualitative character, if any, a man's mental states have. But what usually underlies this is the idea that the man himself has a more 'direct' access to this qualitative character than behavior can possibly provide, namely introspection. And introspection, whatever else it is, is the link between a man's mental states and his beliefs about (or his knowledge or awareness of) those states. So one way of putting our question is to ask whether anything could be evidence (for anyone) that someone was not in pain, given that it follows from the states he is in, plus the psychological laws that are true of him (the laws which describe the relationships of his states to one another and to input and output), that the totality of possible behavioral evidence *plus* the totality of possible introspective evidence points unambiguously to the conclusion that he is in pain? I do not see how anything could be. To be sure, we can imagine (perhaps) that 'cerebroscopes' reveal that the person is not in some neurophysiological state that we ourselves are always in when we are (so we think) in pain. But this simply raises the question, on what basis can we say that *we* have genuine pain (i.e., a state having a qualitative character as well as playing the appropriate functional role in its relationships to input, output, and other psychological states)? Here it seems that if the behavioral and introspective evidence are not enough, nothing could be enough. But if they are enough in the case of us, they are enough in the case of our hypothetical man. In any event, if we are given that a man's state is functionally identical with a state that in us is pain, it is hard to see how a physiological difference between him and us could be any evidence at all that his states lack qualitative character; for if anything can be evidence for us about his psychological state, the evidence that his state is functionally equivalent to ours is *ipso facto* evidence that any physiological difference between us and him is irrelevant to whether, although not to how, the state of pain is realized in him.

To hold that it is logically possible (or, worse, nomologically possible) that a state lacking qualitative character should be functionally identical to a state having qualitative character is to make qualitative character irrelevant both to what we can take ourselves to know in knowing about the mental states of others and also to what we can take ourselves to

know in knowing about our own mental states. There could (on this view) be no possible physical effects of any state from which we could argue by an 'inference to the best explanation' that it has qualitative character; for if there were, we could give at least a partial functional characterization of the having of qualitative character by saying that it tends to give rise, in such and such circumstances, to those physical effects, and could not allow that a state lacking qualitative character could be functionally identical to a state having it. And for reasons already given, if cases of 'absent qualia' were possible, qualitative character would be necessarily inaccessible to introspection. If qualitative character were something that is irrelevant in this way to all knowledge of minds, self-knowledge as well as knowledge of others, it would not be at all 'unacceptable', but would instead be just good sense, to deny that pains must have qualitative character. But of course it is absurd to suppose that ordinary people are talking about something that is in principle unknowable by anyone when they talk about how they feel, or about how things look, smell, sound, etc. to them. (Indeed, just as a causal theory of knowledge would imply that states or features that are independent of the causal powers of the things they characterize would be in principle unknowable, so a causal theory of reference would imply that such states and features are in principle unnamable and inaccessible to reference.) And if, to return to sanity, we take qualitative character to be something that can be known in the ways we take human feelings to be knowable (at a minimum, if it can be known introspectively), then it is not possible, not even logically possible, for a state that lacks qualitative character to be functionally identical to a state that has it.

This is not a 'verificationist' argument. It does not assume any general connection between meaningfulness and verifiability (or knowability). What it does assume is that if there is to be any reason for supposing (as the 'absent qualia argument' does) that it is essential to pain and other mental states that they have 'qualitative character', then we must take 'qualitative character' to refer to something which is knowable in at least some of the ways in which we take pains (our own and those of others) to be knowable. It also assumes that if there could be a feature of some mental state that was entirely independent of the causal powers of the state (i.e., was such that its presence or absence would make no difference to the state's tendencies to bring about other states, and so

forth), and so was irrelevant to its 'functional identity', then such a feature would be totally unknowable (if you like, this assumes a causal theory of knowledge).

Against this argument, as against an earlier one, it may be objected that the other psychological states by relation to which (along with inputs and outputs) a given psychological state is functionally defined must not include any states that cannot themselves be functionally defined. For, it may be said, the states I have called 'qualitative beliefs' can no more be functionally defined than can qualitative states themselves. The most important relationship of these states to other states would appear to be their relationship to the qualitative states that characteristically give rise to them, yet (so the argument goes) the latter cannot be functionally defined and so cannot legitimately be referred to in functional definitions of the former. Moreover (remembering that the possibility of cases of *inverted* qualia is not here being questioned), it seems plausible to suppose that if two people differed in the qualitative character of their pains, but in such a way that the difference would not be revealed in any possible behavior, then they would also differ in their qualitative beliefs, and this difference too would be such that its existence could not be revealed in any possible behavior. And if this is possible, there seems as much reason to deny that qualitative beliefs are capable of functional definition as there is to deny that qualitative states are capable of functional definition.

This objection does not touch one important point implicit in my argument, namely that we cannot deny, without being committed to an intolerable skepticism about the pains of others, that someone's saying that he feels a sharp pain is good evidence that he has some qualitative state or other, and is so because someone's saying this is, normally, an *effect* of his having a state having qualitative character – and this by itself strongly suggests that if a mental state of one person has qualitative character, and an otherwise similar state of another person lacks qualitative character, then the states differ in the ways they tend to influence behavior ('output') and hence differ functionally. Still, the possibility of 'inverted qualia' does seem to imply that qualitative states, and hence qualitative beliefs, cannot be functionally defined. To see whether this is compatible with functionalism, and whether it undercuts the argument given above, we need to consider in what sense it is true that quali-

tative states (and qualitative beliefs) are not functionally definable, and what limits there are on the ways in which reference to mental states that are not functionally definable can enter into functional definitions of other mental states.

In order to consider these questions I wish to change examples, and shift our consideration from the case of pain to that of visual experience. There are two reasons why such a shift is desirable. First, the possibility of 'spectrum inversion' (one person's experience of colors differing systematically, in its qualitative or phenomenological character, from another person's experience of the same colors) seems to me far less problematical than the possibility of 'qualia inversion' in the case of pain (pain feeling radically different to different persons). Second, and related to this, it is much easier to distinguish seeing blue (for example) from its qualitative character than it is to distinguish pain from its qualitative character, and accordingly much easier to consider how reference to qualitative states might enter into a functional account of seeing colors than it is to consider how reference to such states might enter into a functional account of pain.

4. If I see something, it looks somehow to me, and the way it looks resembles and differs, in varying degrees and various respects, the ways other things look to me or have looked to me on other occasions. It is because similarities and differences between these 'ways of being appeared to' correlate in systematic ways with similarities and differences between objects we see that we are able to see these objects and the properties of them in virtue of which the similarities and differences obtain.[13] Being appeared to in a certain way, e.g., things looking to one the way things now look to me as I stare out my window, I take to be a qualitative state. So seeing essentially involves the occurrence of qualitative states. Moreover, reference to these qualitative states enters into what looks very much like a functional account of seeing. For it would seem that what it means to say that someone sees something to be blue is something like the following:

S sees something to be blue if and only if
(1) *S* has a repertory of qualitative states which includes a set of states *K* which are associated with the colors of objects in such a way that (a) visual stimulation by an object of a certain color under 'standard conditions' produces in the person the

associated qualitative state, and (b) the degrees of 'qualitative' or 'phenomenological' similarity between the states in K correspond to the degrees of similarity between the associated colors, and (2) person S (a) is at present in the qualitative state associated with the color blue, (b) is so as the result of visual stimulation by something blue and (c) believes, because of (a) and (b), that there is something blue before him.[14]

I must now qualify the assertion that 'being appeared to' in a certain way is a qualitative state. If asked to describe how he is appeared to, or, more naturally, how things look to him, a man might say, among other things, that a certain object looks blue to him, or that it looks to him as if he were seeing something blue, or (if he is a philosopher who speaks the 'language of appearing') that he is 'appeared-blue-to'. And it is natural to make it a condition of someone's being appeared-blue-to that he be in the qualitative state that is, in him at that time, associated with visual stimulation by blue things; that is, it is natural to give an analysis of 'S is appeared-blue-to' which is the same as the above analysis of 'S sees something to be blue' except that clauses (b) and (c) of condition (2) are deleted. But if we do this, then being appeared-blue-to will not itself be a qualitative state. Or at any rate, this will be so if spectrum inversion is possible. We might sum up the situation by saying that being appeared-blue-to is, on the proposed analysis, a functional state whose functional characterization requires it always to have some qualitative character (or other) but does not require it to have the same qualitative character in different persons (assuming the possibility of intersubjective spectrum inversion) or in the same person at different times (assuming the possibility of intrasubjective spectrum inversion). But this raises again the question of whether qualitative states are themselves functionally definable and, if they are not, whether they can legitimately be referred to in functional characterizations of other mental states.

The expression 'appeared-blue-to' could, I think, have a use in which it would stand for a qualitative state. I could 'fix the reference' of this expression by stipulating that it refers to (or, since it is a predicate rather than a singular term, that it predicates or ascribes) that qualitative state which is at the present time (April, 1974) associated in me with the seeing of blue things.[15] Understanding the expression in this way, if I underwent spectrum inversion tomorrow it would cease to be the case that I am normally appeared-blue-to when I see blue things, and might become

the case that I am normally appeared-yellow-to on such occasions.[16] (By contrast, in the 'functional' sense of 'appeared-blue-to' sketched above, it could be true before and after intrasubjective spectrum inversion that I am normally appeared-blue-to when I see blue things, although of course being appeared-blue-to would have the qualitative character at the later time which another visual state, say, being appeared-yellow-to, had at the earlier time.) I do not think that there would be much utility in having expressions that were, in this way, 'rigid designators' (or 'rigid predicators') of visual qualia. On the other hand, I see no reason in principle why we could not have them. But if we did have them, they could not be functionally defined. Such terms would have to be introduced by Kripkean 'reference fixing' or (what is a special case of this) ostensive definition. To be sure, there is the theoretical possibility of giving a verbal definition of one of these expressions by making use of other expressions of the same sort; just as I might define 'blue' by means of a description of the form 'the color that is not yellow, or red, or green... etc.', so I might define 'being appeared-blue-to' as equivalent to a description of the form 'the color qualia which is neither being appeared-yellow-to, nor being appeared-red-to, nor being appeared-green-to, ... etc.' But this is of very little interest, since it is obviously impossible that names (or predicates) for all visual qualia should be defined in this way without circularity. So, assuming that talk of defining functional states is equivalent to talk of defining names or 'rigid designators' for qualitative states, there seems to be a good sense in which qualitative states cannot be functionally defined.

But what seems to force us to this conclusion is the seeming possibility of spectrum inversion. I think that what (if anything) forces us to admit the possibility of spectrum inversion is the seeming conceivability and detectability of *intra*subjective spectrum inversion. And if we reflect on the latter, we will see, I believe, that while we cannot functionally define particular qualitative states, there is a sense in which we can functionally define the *class* of qualitative states – we can functionally define the identity conditions for members of this class, for we can functionally define the relationships of qualitative (phenomenological) similarity and difference. This is what I shall argue in the following section.

5. Taken one way, the claim that spectrum inversion is possible implies

a claim that may, for all I know, be empirically false, namely that there is a way of mapping determinate shades of color onto determinate shades of color which is such that (1) every determinate shade (including 'muddy' and unsaturated colors as well as the pure spectral colors) is mapped onto some determinate shade, (2) at least some of the shades are mapped onto shades other than themselves, (3) the mapping preserves, for any normally sighted person, all of the 'distance' and 'betweenness' relationships between the colors (so that if shades *a*, *b* and *c* are mapped onto shades *d*, *e* and *f*, respectively, then a normally sighted person will make the same judgments of comparative similarity about *a* in relation to *b* and *c* as about *d* in relation to *e* and *f*), and (4) the mapping preserves all of our intuitions, except those that are empirically conditioned by knowledge of the mixing properties of pigments and the like, about which shades are 'pure' colors and which have other colors 'in' them (so that, for example, if shades *a* and *b* are mapped onto shades of orange and red, respectively, we will be inclined to say that *a* is less pure than *b* and perhaps that it has *b* in it). But even if our color experience is not in fact such that a mapping of this sort is possible, it seems to me conceivable that it might have been – and that is what matters for our present philosophical purposes. For example, I think we know well enough what it would be like to see the world nonchromatically, i.e., in black, white, and the various shades of grey – for we frequently do see it in this way in photographs, moving pictures, and television. And there is an obvious mapping of the nonchromatic shades onto each other which satisfies the conditions for inversion. In the discussion that follows I shall assume, for convenience, that such a mapping is possible for the full range of colors – but I do not think that anything essential turns on whether this assumption is correct.

Supposing that there is such a mapping (and, a further assumption of convenience, that there is only one), let us call the shade onto which each shade is mapped the 'inverse' of that shade. We will have *inter*subjective spectrum inversion if the way each shade of color looks to one person is the way its inverse looks to another person, or, in other words, if for each shade of color the qualitative state associated in one person with the seeing of that shade is associated in another person with the seeing of the inverse of that shade. And we will have *intra*subjective spectrum inversion if there is a change in the way the various shades

of color look to someone, each coming to look the way its inverse previously looked.

What strikes us most about spectrum inversion is that if it can occur intersubjectively there would appear to be no way of telling whether the color experience of two persons is the same or whether their color spectrums are inverted relative to each other. The systematic difference between experiences in which intersubjective spectrum inversion would consist would of course not be open to anyone's introspection. And there would appear to be no way in which these differences could manifest themselves in behavior – the hypothesis that your spectrum is inverted relative to mine and the hypothesis that our color experience is the same seem to give rise to the same predictions about our behavior. Here, of course, we have in mind the hypothetical case in which the various colors have always looked one way to one person and a different way to another person. And the situation seems very different when we consider the case of *intra*subjective spectrum inversion. In the first place, it seems that such a change would reveal itself to the introspection, or introspection *cum* memory, of the person in whom it occurred. But if this is so, other persons could learn of it through that person's reports. Moreover, and this is less often noticed, there is non-verbal behavior, as well as verbal behavior, that could indicate such a change. If an animal has been trained to respond in specific ways to objects of certain colors, and then begins, spontaneously, to respond in those ways to things of the inverse colors, and if it shows surprise that its responses are no longer rewarded in the accustomed ways, this will surely be some evidence that it has undergone spectrum inversion. In the case of a person we could have a combination of this sort of evidence and the evidence of the person's testimony.[17]

If we did not think that we could have these kinds of evidence of intrasubjective spectrum inversion, I think we would have no reason at all for thinking that spectrum inversion of any sort, intrasubjective or intersubjective, is even logically possible. To claim that spectrum inversion is possible but that it is undetectable even in the intrasubjective case would be to sever the connection we suppose to hold between qualitative states and introspective awareness of them (between them and the qualitative beliefs to which they give rise), and also their connections to perceptual beliefs about the world and, *via* these beliefs, to behavior.

No doubt one could so *define* the term 'qualitative state' as to make it inessential to qualitative states that they have these sorts of connections. But then it would not be in virtue of similarities and differences between 'qualitative states' (in that defined sense) that things look similar and different to people, and the hypothesis that people differ radically in what 'qualitative states' they have when they see things of various colors would be of no philosophical interest, and would not be the 'inverted spectrum hypothesis' as usually understood. Indeed, the supposition that intrasubjective spectrum inversion could occur, but would be undetectable, is incoherent in much the same way as the 'absent qualia hypothesis', i.e., the supposition that states 'functionally identical' to states having qualitative content might themselves lack qualitative content. Neither supposition makes sense unless the crucial notions in them are implicitly defined, or redefined, so as to make the supposition empty or uninteresting.

But what, then, are we supposing about qualitative states, and about the relationships of qualitative or phenomenological similarity and difference between these states, in supposing that intrapersonal spectrum inversion *would* be detectable? In what follows I shall speak of token qualitative states as 'experiences', and will say that experiences are 'co-conscious' if they are conscious to a person at the same time, where an experience counts as conscious to a person when he correctly remembers it as well as when he is actually having it. One thing we are supposing, if we take intrasubjective spectrum inversion to be detectable in the ways I have indicated, is that when experiences are co-conscious the similarities between them tend to give rise to belief in the existence of objective similarities in the physical world, namely similarities between objects in whose perception the experiences occurred, and differences between them tend to give rise to belief in the existence of objective differences in the world. And these beliefs, in turn, give rise (in combination with the person's wants and other mental states) to overt behavior which is appropriate to them. This explains how there can be non-verbal behavior that is evidence of spectrum inversion; the behavior will be the manifestation of mistaken beliefs about things which result from the fact that, in cases of intrasubjective spectrum inversion, things of the same color will produce qualitatively different experiences after the inversion than they did before, while things of each color will produce, after the inver-

sion, experiences qualitatively like those produced by things of a different color before the inversion.

But even if, for some reason, a victim of spectrum inversion were not led to have and act on mistaken beliefs about objective similarities and dissimilarities in this way, we could still have evidence that his spectrum had inverted – for he could tell us that it had. And in supposing that *he* can know of the spectrum inversion in such a case, and so be in a position to inform us of it, we are supposing something further about the relationships of qualitative similarity and difference, namely that when they hold between co-conscious experiences, this tends to give rise to introspective awareness of the holding of these very relationships, i.e., it tends to give rise to correct "qualitative beliefs" to the effect that these relationships hold.

Philosophers who talk of mental states as having behavioral 'criteria' have sometimes said that the criterion of experiences being similar is their subject's sincerely reporting, or being disposed to report, that they are. If we recast this view in functionalist terms, it comes out as the view that what constitutes experiences being qualitatively similar is, in part anyhow, that they give rise, or tend to give rise, to their subject's having a qualitative belief to the effect that such a similarity holds, and, in virtue of this belief, a disposition to make verbal reports to this effect. But as a functional *definition* of qualitative similarity this would of course be circular. If we are trying to explain what it means for experiences to be similar, we cannot take as already understood, and as available for use in our explanation, the notion of believing experiences to be similar.

But no such circularity would be involved in functionally defining the notions of qualitative similarity and difference in terms of the first sort of relationship I mentioned, namely between, on the one hand, a person's experiences being qualitatively similar or different in certain ways, and, on the other, his believing in the existence of certain sorts of objective similarities or differences in the world, and, ultimately, his behaving in certain ways. I believe that a case can be made, although I shall not attempt to make it here, for saying that the tendency of sensory experiences to give rise to introspective awareness of themselves, and of their similarities and differences, is, for creatures having the conceptual capacities of humans, an inevitable by-product of their tendency to give rise to perceptual awareness of objects in the world, and of similarities

and differences between these objects. And my suggestion is that what makes a relationship between experiences the relationship of qualitative (phenomenological) similarity is precisely its playing a certain 'functional' role in the perceptual awareness of objective similarities, namely its tending to produce perceptual beliefs to the effect that such similarities hold. Likewise, what makes a relationship between experiences the relationship of qualitative difference is its playing a corresponding role in the perceptual awareness of objective differences.

This suggestion is, of course, vague and sketchy. But all that I have to maintain here is that the claim that we can give a functional account of qualitative similarity and difference along these lines is no less plausible than the claim that such mental states as belief and desire can be functionally defined. For my aim is not the ambitious one of showing that functionalism provides a fully satisfactory philosophy of mind; it is the much more modest one of showing that the fact that some mental states have 'qualitative character' need not pose any special difficulties for a functionalist. And an important step toward showing the latter is to show that the notions of qualitative similarity and difference are as plausible candidates for functional definition as other mental notions. I conceded earlier that there is a sense in which particular qualitative states cannot be functionally defined. But it will be remembered that what distinguishes qualitative states from other sorts of mental states is that their 'type-identity conditions' are to be given in terms of the notion of qualitative similarity. At the beginning of our discussion, specifying identity conditions in such terms seemed to contrast sharply with specifying them in functional terms. But this contrast becomes blurred if, as I have suggested, the notion of qualitative similarity can itself be defined in functional terms. And if the latter is so, and hence the identity conditions for qualitative states can be specified in functional terms, it seems not inappropriate to say, as I did earlier, that while particular qualitative states cannot be functionally defined, the *class* of qualitative states can be functionally defined.

6. Now let us return to the question of whether it is legitimate to make reference to qualitative states in giving functional definitions of other sorts of mental states.

On one construal of it, functionalism in the philosophy of mind is the

doctrine that mental, or psychological, terms are, in principle, eliminable in a certain way. If, to simplify matters, we take our mental vocabulary to consist of names for mental states and relationships (rather than predicates ascribing such states and relationships), the claim will be that these names can be treated as synonymous with definite descriptions, each such description being formulable, in principle, without the use of any of the mental vocabulary. Mental states will indeed be quantified over, and in some cases identifyingly referred to, in these definite descriptions; but when they are, they will be characterized and identified, not in explicitly mentalistic terms, but in terms of their causal and other 'topic neutral' relations to one another and to physical inputs and outputs.[18]

Now what I have already said implies that names of qualitative states (if we had them) could not be defined as equivalent to such definite descriptions – on the assumption, of course, that 'qualia inversion' is possible. If the causal role played by a given qualitative state (in conjunction with other mental states) in mediating connections between input and output could be played by another qualitative state, and if that qualitative state could play a different role, then it is not essential to the state that it plays that causal role and it cannot be part of the meaning, or sense, of a term that rigidly designates it that the state so designated is *the* state that plays such a causal role. Moreover, since such a term could not be eliminated in this way in favor of a definite description, it could not occur within the definite description which functionally defines the name of some other mental state – assuming that the aim of such functionalist definitions is to eliminate mental terminology in favor of physical and topic neutral terminology.

But there is nothing in this to imply that qualitative states cannot be among the states quantified over in the definite descriptions that define other sorts of mental states. And it seems that it would be quantification over such states, rather than reference to particular states of this kind, that would be needed in the defining of other mental states. If spectrum inversion is possible, we do not want to make the occurrence of any particular qualitative state a necessary condition of seeing (or seeming to see) something blue, but we do want to require that at any given time in the history of a person there is some qualitative state or other that is (at that time) standardly involved in his seeing (or seeming

to see) blue things. The specification of the roles of the qualitative states in the seeing of blue things will no doubt invoke the notions of qualitative similarity and difference; but this causes no difficulties for a functionalist if, as I have suggested, these notions can themselves be functionally defined.

There would appear, however, to be some mental states (other than qualitative states) that cannot be functionally defined in the strong sense here under consideration, namely in such a way that there is no essential (uneliminable) use of mental terminology in the *definiens*. For consider the states I have called 'qualitative beliefs', i.e., beliefs about qualitative states and in particular beliefs to the effect that one is in a particular qualitative state. Qualitative beliefs can be divided into two groups, those in whose propositional content there is reference to particular qualitative states, and those in whose propositional content there is quantification over qualitative states but no reference to particular qualitative states. So far as I can see, qualitative beliefs of the second sort provide no special difficulties for the functionalist; if other sorts of beliefs can be functionally defined, so can these. But qualitative beliefs of the first sort do seem to resist functional definition. Consider the belief I would express if I said 'I am in the state of being appeared-blue-to', using the phrase 'state of being appeared-blue-to' to rigidly designate a particular qualitative state. If we tried to characterize this state of believing functionally, i.e., in terms of its relationships to other mental states and to input and output, it would seem that we would have to make reference in our characterization to the qualitative state the belief is about – we would have to say that the state of believing that one is appeared-blue-to is typically the result of the state of being appeared-blue-to. If so, it is impossible to define such states (qualitative beliefs of the first sort) without making essential use of mental terms.

But this constitutes no obstacle to our functionally defining other sorts of mental states. For while we may want to include in our functional characterizations of some kinds of mental states that they give rise to qualitative beliefs of the first sort (i.e., those in whose propositional content there is reference to particular qualitative states), this need not involve our making identifying reference to beliefs of this sort in our functional characterizations; all that this need involve is quantifying over such beliefs. Thus, for example, we can build it into our functional char-

acterization of pain that being in pain typically results in some qualitative belief to the effect that one has some specific qualitative state, without saying of any specific qualitative state that being in pain tends to give rise to a belief about it. And if quantifying over qualitative states is permissible in giving functional definitions, I see no reason why quantifying over functional beliefs should not be permissible as well.

Now let us return briefly to my argument in section 3 against the possibility of cases of 'absent qualia'. In that argument I pointed out that it is characteristic of pains to give rise to introspective awareness of themselves as having particular qualitative characters, and so to give rise to 'qualitative beliefs', and I used this to argue that any state functionally identical to a state having qualitative character (e.g., a pain) must itself have qualitative character. The objection was raised to this argument that since qualitative beliefs, like qualitative states, cannot be functionally defined, they cannot legitimately enter into a functional account of the 'type-identity conditions' for other mental states. We can now answer this objection. No doubt pains give rise to qualitative beliefs of the sort that (so I am allowing) cannot be functionally defined, i.e., beliefs to the effect that one is having some specific qualitative state. But they also give rise to beliefs to the effect that one is in pain – and if (as the 'absent qualia argument' apparently assumes) pain is necessarily a state having qualitative character, then the belief that one is in pain presumably involves (at least in the case of a reflective person) the belief that one is in a state having some qualitative state or other. And while the latter belief is a qualitative belief, its propositional content quantifies over qualitative states rather than involving reference to particular qualitative states. No reason has been given why qualitative beliefs of this sort should not be regarded as functionally definable. And if they are functionally definable, there is no reason why the tendency of other states to give rise to such beliefs should not be part of what constitutes the functional identity of those other states. And this is all the argument of section 3 requires.

7. Over the last few decades, much of the controversy in the philosophy of mind has involved a battle between two seemingly conflicting sets of intuitions. On the one hand there is the intuition that mental states are somehow logically, or conceptually, connected with physical

states of affairs, in particular the behaviors that are taken to manifest them. This intuition has found expression in a succession of different philosophical positions – logical behaviorism, the 'criteriological' views inspired by Wittgenstein, and, most recently, functional or causal analyses of mental states (these usually being combined with some form of materialism or physicalism).[19] On the other hand there is the intuition that connections between mental states and behavior are, at bottom, contingent; that under the most 'intrinsic' descriptions of mental states, it is a contingent fact that they are related as they are to behavior and to other sorts of physical states. And a common expression of this view has been the claim that spectrum inversion and other sorts of 'qualia inversion' are logically possible; for to say that these are logically possible is apparently to say that what intrinsic, internal character these mental states have, their 'qualitative content', is logically irrelevant to their being related as they are to their bodily causes and behavioral manifestations. I have conceded that there is a substantial element of truth in this view. For I have allowed that spectrum inversion is a possibility, and have allowed that this implies that at least some qualitative states (and qualitative beliefs) cannot be functionally defined. But I believe that there is a substantial element of truth in the other view as well. I think that where the other view – the view that mental states are 'logically' or 'conceptually' connected with behavior – has its greatest plausibility is in its application to such states as desire and belief, and I think that these states do not have 'qualitative character' in the sense that here concerns us, although they may sometimes be accompanied by qualitative states. But as I have tried to show, I think that even qualitative states can be accommodated within the framework of a functional, or causal, analysis of mental states. While it may be of the essence of qualitative states that they are 'ineffable' in the sense that one cannot say in general terms, or at any rate in general terms that do not include names of qualitative states, what it is for a person to be in a particular qualitative state, this does not prevent us from giving a functional account of what it is for a state to be a qualitative state, and of what the identity conditions for qualitative states are. Thus it may be possible to reconcile these firmly entrenched, and seemingly conflicting, intuitions about the contingency or otherwise of relations between mental states and the physical world.

There are a number of issues that would have to be investigated before it could be claimed that this attempted reconciliation is successful. The account of qualitative similarity and difference that I have suggested was tailored to the case of perceptual experiences, and it needs to be considered whether it can be plausibly applied to sensations like pains. What its application to the case of pain may require is the acceptance of the view of pains as somatic sense impressions, i.e., impressions (which need not be veridical) of bodily injuries and the like.[20] Also, this account of qualitative similarity and difference is tailored to the case in which the experiences being compared are experiences of one and the same person, and it needs to be considered whether it gives sense, and the right sort of sense, to intersubjective comparisons of experiences. This would involve, among other things, a consideration of whether it is possible for experiences of different persons to be 'co-conscious' in the sense defined earlier; and I think this reduces to the question of whether it is possible for there to be 'fusion' between persons of the sort envisaged in some recent discussions of personal identity, i.e., a merging of two persons into a single person (or single subject of consciousness) who then remembers, and is able to compare, the experiences the persons had prior to the fusion. (It is worth noting that if fusion is possible, then it is not after all the case that no possible behavior would reveal whether the color experience of two persons was the same or whether their color spectrums were inverted relative to each other; for were the persons to fuse, the behavior of the resulting person could presumably settle this question.) But these are all complex issues, and I shall not attempt to discuss them here.[21]

Cornell University

NOTES

[1] N. J. Block and J. A. Fodor, 'What Psychological States are Not', *The Philosophical Review* LXXXI (1972), p. 165.
[2] *Op. cit.*, p. 173.
[3] *Op. cit.*, pp. 173–174.
[4] *Op. cit.*, p. 172.
[5] *Op. cit.*, p. 173.
[6] *Op. cit.*, p. 172. It is worth noting that this assumption, or one very much like it, plays a crucial role in Saul Kripke's recent arguments against the psychophysical identity theory; Kripke expresses it by saying that pain "is not picked out by one

of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality. Thus pain... is not rigidly designated by 'pain' but the reference of the designator is determined by an essential property of the referent" ('Naming and Necessity', in D. Davidson and H. Harman (eds.) *Semantics of Natural Language* (D. Reidel Publ. Co., Dordrecht-Holland, 1972, p. 340).

[7] Block and Fodor mention another way, besides that mentioned in the text, in which a functionalist might try to meet the inverted qualia argument; he might maintain that "though inverted qualia, *if they occurred*, would provide counterexamples to his theory, as a matter of nomological fact it is impossible that functionally identical psychological states should be qualitatively distinct" (p. 172). The thought here must be that the mere logical, or conceptual, possibility of qualia inversion is not incompatible with functionalism. It would seem, however, that if the actual occurrence of inverted qualia would provide counterexamples to functionalism (as the envisioned reply concedes). then the mere logical possibility of inverted qualia is incompatible with functionalism; pain cannot be *identical* with a given functional state if there is a possible world, even a logically but not nomologically possible world, in which the functional state exists without pain existing, or *vice versa*. (On the general claim about identity here being invoked, namely that if *a* and *b* are identical they must be identical in any logically possible world in which either exists, see Kripke's 'Naming and Necessity', already cited, and his 'Identity and Necessity', in Milton K. Munitz, (ed.), *Identity and Individuation*, New York, 1971.

[8] Block and Fodor, *op. cit.*, p. 173.

[9] Don Locke, *Myself and Others*, Oxford, 1968, p. 101.

[10] Alan Donagan, 'Wittgenstein on Sensations,' in G. Pitcher (ed.), *Wittgenstein: The Philosophical Investigations*, Garden City, New York, 1966.

[11] Block and Fodor, *loc. cit.*

[12] Just what is the relationship that a state must have to a qualitative state in order to have the qualitative character corresponding to that state? It cannot be, in the cases that concern us, the relationship of identity (that would permit only qualitative states to have qualitative character, and would not permit us to speak of the qualitative character of states whose 'type-identity' conditions are given in functional terms). And presumably it must be something stronger than the relationship 'is accompanied by', or 'is coinstantiated with'. The best I can do is to say that a particular token of a state *S* had the qualitative character corresponding to qualitative state *Q* if on the occasion in question the tokening (instantiation) of *S* essentially involved the tokening (instantiation) of *Q*. Possibly, but I am not sure of this, we could strengthen this, and make it less vague, by saying that on such occasions the token of *S is* a token of *Q*.

[13] The 'being appeared to' terminology I take from Roderick Chisholm; see his "'Appear', 'Take', and 'Evident'", in R. J. Swartz (ed.), *Perceiving, Sensing and Knowing*, Garden City, New York, 1965, especially p. 480, footnote 6. One is 'appeared to' both in cases of veridical perception and in cases of illusion and hallucination, and can be appeared to in the same ways in all of these sorts of cases. The technical locution 'appeared-blue-to' is used in the text as an abbreviation for the locution 'sees or seems to see something blue' (on a 'nonepistemic' understanding of that locution).

[14] As an analysis this will not quite do. I can see something to be blue even though it looks green (i.e., even if my visual qualitative state is that associated with green), if I have been 'tipped off' that in these circumstances blue things look green.

[15] I take the notion of 'reference fixing', and the notion of a 'rigid designator' employed below, from Saul Kripke; see his 'Naming and Necessity', pp. 269–275 and *passim*. The use of a definite description 'the x such that Fx' to 'fix the reference' of a term T contrasts with defining T as equivalent in meaning to, i.e., as an abbreviation of, the definite description; in the former case, but not in the latter, the statement 'if T exists, then T is the x such that Fx' will be contingently rather than necessarily true. An expression is a rigid designator if it designates the same object in all possible worlds (or in all possible worlds in which it designates anything). According to Kripke, ordinary names are rigid designators, while definite descriptions are not. When a definite description is used to introduce a name (and hence a rigid designator), it is used to 'fix its reference' rather than to 'define' it or give its 'meaning'.

[16] My distinction between the 'functional' sense of 'appeared-blue-to' and a (possible) sense in which it rigidly designates (or, better, rigidly predicates) a qualitative state is similar to Chisholm's distinction between the 'comparative' and 'noncomparative' senses of expressions like 'looks blue'. See his *Perceiving: A Philosophical Study*, Ithaca, 1957, Chapter Four.

[17] Sometimes it is suggested that if someone reported having undergone spectrum inversion, the most reasonable thing for us to conclude would be that something had gone awry with his grasp of the color vocabulary. This overlooks the fact that such a report could be backed up by behavioral evidence of a non-verbal sort. And I think we can imagine a series of events that would leave us no alternative but to conclude that spectrum inversion had occurred. Let us represent the color spectrum by a vertical line, and let us, arbitrarily, divide the line into six equal segments, labeling these from top to bottom with the first six letters of the alphabet. And now consider the case of George. At time t_1 George's color experience, and his use of color words, was perfectly normal. But at time t_2 he tells us that a remarkable change has occurred; while most things look to him just as they used to, or look different only in ways that might be expected (e.g., if there is painting being done), a sizable minority of objects look to him very different than they did before, and he knows, from consulting other persons and from spectroscopic evidence, that in fact these objects have not undergone any significant change in color. George describes the change by saying that if he now looks at what we would regard as a normal spectrum, it looks the way a spectrum would have looked at t_1 if the end segments, A and F, had been interchanged and rotated one hundred and eighty degrees, the positions of the other segments remaining unchanged. According to this, the structure of George's color experience at t_2 is different from its structure at t_1. And since the putative change involves a change in structure, our evidence that it occurred need not be limited to George's testimony. George's claim will be supported by his recognitional and discriminatory behavior if, as we will suppose, he finds it easy to discriminate certain shades of color, for example those on either side of the boundary between segments A and B of the spectrum, which he formerly found it difficult to discriminate (and which the rest of us still find it difficult to discriminate), and sometimes finds it difficult to discriminate between different shades, for example if one is near the bottom boundary of segment A and the other is near the bottom boundary of segment E, which he formerly found it easy to discriminate (and which the rest of us still find it easy to discriminate). To continue the story, at time t_3 George tells us that another such change has occurred and added itself, as it were, to the first one; this time it is as if segments B and E of the spectrum had been interchanged and rotated. Again we can suppose that there is behavioral evidence to substantiate his claim. Finally,

at time t_4 he tells us that still another such change has occurred; this time it is as if segments *C* and *D* had been interchanged and rotated. And again there is the substantiating behavioral evidence. But at t_4, unlike at t_2 and t_3, George's judgments of color similarity and difference will coincide with ours and with those he made at t_1 (allowing, of course, for whatever objective changes in color may have occurred in the interim); at t_4 the 'structure' of George's color experience will be the same as it was at t_1. Yet George reports that his color experience is systematically different from what it was at t_1; each color looks the way its inverse looked previously. And this claim of George's seems to be supported by the behavioral evidence that supported his claims that there were changes in his color experience between t_1 and t_2, between t_2 and t_3, and between t_3 and t_4; for these partial changes add up to a total spectrum inversion.

[18] This account of what functional definition would amount to, and the elaboration of it that follows, is based loosely on David Lewis' account in 'Psychophysical and Theoretical Identification', *Australasian Journal of Philosophy* **50**, (December, 1972), pp. 249–258.

Starting with the 'theory' which consists of the set of 'platitudes' about relations of mental states to one another and to input and output which it is plausible to regard as analytic or quasi-analytic, we can define the mental terms in that theory (supposing them, for simplicity, to be names of mental states) in the following way. We first write the theory as a single conjunctive sentence. We then replace each of the mental terms in the theory with a different variable, forming an open sentence. We then prefix quantifiers which transform the open sentence into the 'modified Ramsey sentence' of the theory, which says (in effect) that there exists a unique n-tuplet of states satisfying the open sentence. We are now in a position to define any of the mental terms that occurred in the original theory. Supposing that T_i is the term we wish to define, and y_i is the variable we replaced it with in forming the modified Ramsey sentence, we can turn the modified Ramsey sentence into a definite description by (1) adding to the open sentence within the scope of the initial quantifiers the conjunct '$y_i=x$,' where 'x' is a variable that does not occur in the modified Ramsey sentence, and (2) prefixing the whole sentence with a definite description operator binding 'x'. What we then get is something of the form: $(\iota x)\ (\exists y_1)\ldots(\exists y_i)\ldots(\exists y_n)\ (\ldots\ldots y_i\ldots\ldots\ \&\ y_i=x)$. In this description there will occur no mental terms. And we can define T_i as being synonymous with this description.

I should emphasize that what I am characterizing here is only one version of functionalism. Many philosophers who would regard themselves as functionalists would disavow any intention of giving, or providing a recipe for giving, any sort of meaning analysis of psychological terms.

[19] Some advocates of causal or functional theories of the mind, especially those who would not accept the characterization of functionalism in section 6 and footnote 18, would object to being put in this company. But others have clearly seen their accounts as incorporating what is correct in, or as explaining the intuitions which make plausible, behavioristic and criteriological views. See, for example, David Lewis, *op. cit.*, p. 257, David Armstrong, *A Materialist Theory of the Mind*, London, 1968, p. 92, and Alvin Goldman, *A Theory of Human Action*, Englewood Cliffs, N.J., 1970, p. 112.

[20] Such a view has in fact been advanced by D. M. Armstrong and by George Pitcher. See Armstrong, *op. cit.*, p. 313ff., and Pitcher's 'Pain Perception', *The Philosophical Review* **LXXIX** (1970), pp. 368–393.

[21] I have benefited from discussions on this topic with Jonathan Bennett and Keith

Lehrer, and am grateful to Bennett, and to N. L. Block, for criticisms of an earlier version of the paper. The paper was written while I was a Fellow at the Center for Advanced Study in the Behavioral Sciences, in Stanford, California, and I would like to express my gratitude to that institution.