Functionalism, Qualia, and Intentionality

PAUL M. CHURCHLAND
PATRICIA SMITH CHURCHLAND
University of Manitoba

Functionalism—construed broadly as the thesis that the essence of our psychological states resides in the abstract causal roles they play in a complex economy of internal states mediating environmental inputs and behavioral outputs—seems to us to be free from any fatal or essential shortcomings. Functionalism-on-the-hoof is another matter. In various thinkers this core thesis is generally embellished with certain riders, interpretations, and methodological lessons drawn therefrom. With some of the more prominent of these articulations we are in some disagreement, and we shall turn to discuss them in the final section of this paper. Our primary concern, however, is to *defend* functionalism from a battery of better-known objections widely believed to pose serious or insurmountable problems even for the core thesis outlined above. In sections I and II we shall try to outline what form functionalism should take in order to escape those objections.

I. *Four Problems Concerning Qualia*

'Qualia' is a philosophers' term of art denoting those intrinsic or monadic properties of our sensations discriminated in introspection. The quale of a sensation is typically contrasted with its causal, relational, or functional features, and herein lies a problem for functionalism. The quale of a given sensation—pain, say—is at best contingently connected with the causal or functional properties of that state; and yet common intuitions insist that said quale is an essential element of pain, on some views, *the* essential element. Functionalism, it is concluded, provides an inadequate account of our mental states.

Before addressing the issues in greater detail, let us be clear about what the functionalist need not deny. He need not and should not deny that our sensations have intrinsic properties, and he should agree as well that those properties are the principal means of our introspective discrimination of one kind of sensation from another. What he is committed to denying is that any particular quale is

essential to the identity of any particular type of mental state. Initially they may seem to be essential, but reflection will reveal that they do not have and should not be conceded that status. In what follows we address four distinct but not unrelated problems. Each problem is manageable on its own, but if they are permitted to band together for a collective assault, the result is rather confusing and formidable, in the fashion of the fabled Musicians of Bremen. With the problems separated, our strategy will be to explain and exploit the insight that intrinsic properties *per se* are no anathema to a functionalist theory of mental states.

A. The Problem of Inverted/Gerrymandered Qualia

This problem is just the most straightforward illustration of the general worry that functionalism leaves out something essential. The recipe for concocting the appropriate intuitions runs thus. Suppose that the sensations having the quale typical of pain in you play the functional role of pleasure sensations in someone else, and the quale typical of pleasure sensations in you are had instead by the sensations that have the functional role of pain in him. Functionally, we are to suppose, the two of you are indistinguishable, but his pleasure/pain qualia are simply inverted relative to their distribution among your own sensations, functionally identified. A variation on the recipe asks us to imagine someone with an inverted distribution of the color qualia that characterize your own visual sensations (functionally identified). He thus has (what you would introspectively identify as) a sensation of red in all and only those circumstances where you have a sensation of green, and so forth.

These cases are indeed imaginable, and the connection between quale and functional syndrome is indeed a contingent one. Whether it is the quale or the functional syndrome that determines type–identity qua psychological state, we must now address. The intuitions evoked above seem to confound functionalist pretensions. The objection to functionalism is that when the inversion victim has that sensation whose functional properties indicate pleasure, *he is in fact feeling pain*, functional properties notwithstanding; and that when the victim of a spectrum inversion says, "I have a sensation of green" in the presence of a green object, *he is in fact having a sensation of red*, functional properties notwithstanding. So far as type–identity of psychological states is concerned, the objection concludes, sameness of qualitative character dominates over sameness of functional role.

Now there is no point in trying to deny the possibilities just outlined. Rather, what the functionalist must argue is that they are

better described as follows. "Your pains have a qualitative character rather different from that of his pains, and your sensations–of–green have a qualitative character rather different from that of his sensations–of–green. Such internal differences among the same psychological states are neither inconceivable, nor even perhaps very unusual." That is to say, the functionalist should concede the juggled qualia, while continuing to reckon type–identity in accordance with functional syndrome. This line has a certain intuitive appeal of its own, though rather less than the opposing story, at least initially. How shall we decide between these competing intuitions? By isolating the considerations that give rise to them, and examining their integrity.

The "pro–qualia" intuitions, we suggest, derive from two main sources. To begin with, all of us have a strong and entirely understandable tendency to think of each type of psychological state as constituting a *natural kind*. After all, these states do play a vigorous explanatory and predictive role in everyday commerce, and the common–sense conceptual framework that comprehends them has all the features and functions of a sophisticated empirical theory (see Wilfrid Sellars, 1956; and Paul Churchland, 1979). To think of pains, for example, as constituting a natural kind is to think of them as sharing an *intrinsic nature* that is common and essential to *every* instance of pain. It is understandable then, that the qualitative character of a sensation, the only non–relational feature to which we have access, should present itself as being that essential element.

Our inclination to such a view is further encouraged by the fact that one's introspective discrimination of a sensation's qualitative character is far and away the most immediate, most automatic, most deeply entrenched, and (in isolation) most authoritative measure of what sensations one has. In one's own case, at least, the functional features of one's sensations play a minor role in one's recognition of them. It is as if one had a special access to the intrinsic nature of any given type of sensation, an access that is independent of the purely contingent and causal features that constitute its functional role.

Taken conjointly, these considerations will fund very strong intuitions in favor of qualia as *the* determinants of type–identity for psychological states. But though natural enough, the rationale is exceptionally feeble on both points.

Take the first. However accustomed or inclined we are to think of our psychological states as constituting natural kinds, it is vital to see that it is not a semantic or conceptual matter, but an objective *empirical* matter, whether or not they do. Either there is an objective

intrinsic nature common to all cases of, e.g., pain, as it occurs in humans, chimpanzees, dogs, snakes, and octopi, or there is not. And the fact is, the functionalist can point to some rather persuasive considerations in support of the view that there is not. Given the physiological and chemical variety we find in the nervous systems of the many animals that feel pain, it appears very unlikely that their pain states have a common physical nature underlying their common functional nature (see Hilary Putnam, 1971). It remains possible that they all have some intrinsic *non*–physical nature in common, but dualism is profoundly implausible on sheer evolutionary grounds. (The evolutionary process just *is* the diachronic articulation of matter and energy. If we accept an evolutionary origin for ourselves, then our special capacities must be construed as the capacities of one particular articulation of matter and energy. This conclusion is confirmed by our increasing understanding of the nervous system, both of its past evolution and its current regulation of behavior.) In sum, the empirical presumption *against* natural–kind status for psychological states is substantial. We should not place much trust, therefore, in intuitions born of an uncritical prejudice to the contrary. Such intuitions may reflect ordinary language more or less faithfully, but they beg the question against functionalism.

The facts of introspection provide no better grounds for thinking sensations to constitute natural kinds, or for reckoning qualia as their constituting essences. That the qualitative character Q of a psychological state S should serve as the standard ground of S's introspective discrimination is entirely consistent with Q's being a non–essential feature of S. The black and yellow stripes of a tiger serve as the standard ground on which tigers are visually discriminated from other big cats, but the stripes are hardly an essential element of tigerhood: there are albino tigers, as well as the very pale Himalayan tigers. The telling question here is this: why should the qualia of our familiar psychological states be thought any different? We learn to pick out those qualia in the first place, from the teeming chaos of our inner lives, only because the states thus discriminated are also the nexus of various generalizations connecting them to other inner states, to environmental circumstances, and to overt behaviors of interest and importance to us. Had our current taxonomy of introspectible qualia been *un*successful in this regard, we would most certainly have thrown it over, centered our attention on different aspects of the teeming chaos within, and recarved it into a different set of similarity

classes—a set that *did* display its objective integrity by its many nomic connections, both internal and external. In short, the internal world comes pre–carved into observational kinds no more than does the external world, and it is evident that the introspective taxonomies into which we eventually settle are no less shaped by considerations of explanatory and causal coherence than are the taxonomies of external observation.

It is therefore a great irony, it seems to us, that anyone should subsequently point to whatever qualia our introspective mechanisms have managed tenuously to fix upon as more–or–less usable indicators of nomologically interesting states, and claim *them* as constituting the *essence* of such states. It is of course distantly possible that our mechanisms of introspective discrimination have lucked onto the constituting essences of our psychological states (assuming, contrary to our earlier discussion, that each type *has* a uniform natural essence), but a priori that seems about as likely as that the visual system lucked onto the constituting essence of tigerhood when it made black–on–yellow stripes salient for distinguishing tigers.

Therefore, it seems very doubtful that the type–identity of any psychological state derives from its sharing in any uniform natural essence. Moreover, even if it does so share, it seems entirely unlikely that introspection provides any special access to that essence. Consequently, this beggars the intuition which sustains the inverted–qualia objections.

The preceding investigation into the weight and significance of factors determining type–identity of psychological states does more than that, however. It also enriches the competing intuition, namely, that the type–identity of psychological states is determined by functional characteristics. To repeat the point made earlier, since the taxonomy of observational qualia constructed by the questing child *follows* the discovered taxonomy of states as determined by interesting causal roles, it is evident that sameness of functional role dominates over differences in qualitative character, so far as the type–identity of psychological states is concerned. That a single category, unified by functional considerations, can embrace diverse and disparate qualitative characters has a ready illustration, ironically enough, in the case of pain.

Consider the wide variety of qualia wilfully lumped together in common practice under the heading of pain. Compare the qualitative character of a severe electric shock with that of a sharp blow to the kneecap; compare the character of hands dully aching

from making too many snowballs with the piercing sensation of a jet engine heard at very close range; compare the character of a frontal headache with the sensation of a scalding pot grasped firmly. It is evident that what unites sensations of such diverse characters is the similarity in their functional roles. The sudden onset of any of them prompts an involuntary withdrawal of some sort. Our reaction to all of them is immediate dislike, and the violence of the dislike increases with the intensity and duration of the sensation. All of them are indicators of physical trauma of some kind, actual or potential. All of them tend to produce shock, impatience, distraction, and vocal reactions of familiar kinds. Plainly, these collected causal features are what unite the class of painful sensations, not some uniform quale, invariant across cases. (For a general account of the intentionality of our sensations, in which qualia also retreat into the background, see Paul Churchland, 1979: ch. 2.)

The converse illustration is formed by states having a uniform or indistinguishable qualitative character, states which are nevertheless distinguished by us according to differences in their functional roles. For example, our emotions have a certain qualitative character, but it is often insufficient to distinguish which emotion should be ascribed. On a particular occasion, the felt knot in one's soul might be mild sorrow, severe disappointment, or gathering despair, and which of these it is—really is—would depend on the circumstances of its production, the rest of one's psychological state, and the consequences to which it tends to give rise. Its type–identity need not be a mystery to its possesser—he has introspective access to some of the context which embeds it—but the identification remains unmakeable on qualitative grounds alone. Similarly, a therapist may be needed, or a thoughtful friend, to help you distinguish your decided unease about some person as your hatred for him, envy of him, or simple fear of him. The felt quality of your unease may be the same for each of these cases, but its causes and effects would be significantly different for each. Here again, functional role is the dominant factor in the type–identity of psychological states.

The reason that functional role dominates introspectible qualitative differences and similarities is not that the collected laws descriptive of a state's functional relations are analytically true, or that they exhaust the essence of the state in question (though withal, they may). The reason is that the common–sense conceptual framework in which our psychological terms are semantically

embedded is an *empirical theory*. As with theoretical terms generally, their changeable position in semantic space is fixed by the set of theoretical laws in which they figure. In the case of folk psychology, those laws express the causal relations that connect psychological states with one another, with environmental circumstances, and with behavior. Such laws need not be seen, at any given stage in our growing understanding, as *exhausting* the essence of the states at issue, but at any given stage they constitute the best–founded and most authoritative criterion available for identifying those states.

We conclude against the view that qualia constitute an essential element in the type–identity of psychological states. Variations within a single type are both conceivable and actual. The imagined cases of qualia inversion are of interest only because they place directly at odds intuitions that normally coincide: the non–inferential impulse of observational habit against the ponderous background of theoretical understanding. However, the qualitative character of a sensation is a relevant mark of its type–identity only insofar as that character is the uniform concomitant of a certain repeatable causal syndrome. In the qualia–inversion thought experiments, that uniformity is broken, and so, in consequence, is the relevance of those qualia for type–identity, at least insofar as they can claim a *uniform* relevance across people and across times.

B. The Problem of Absent Qualia

The preceding arguments may settle the qualia–inversion problem, but the position we have defended is thought to raise in turn an even more serious problem for functionalism (see Ned Block and Jerry Fodor, 1972; and Block, 1978). If the particular quale a sensation has contributes nothing to its type–identity, what of a "psychological" system functionally isomorphic with us, whose functional states have no qualia whatever? Surely such systems are possible (nomically as well as logically), runs the objection. Surely functionalism entails that such a system feels pain, warmth, and so on. But since its functional states have no qualitative character whatever, surely such a system *feels nothing at all*. Functionalism, accordingly, must be false.

This argument is much too glib in the contrast it assumes between functional features (which supposedly matter to functionalism) and qualitative character (which supposedly does not). As the functionalist should be the first to admit, our various sensations are introspectively discriminated by us on the basis of their qualitative

character, and any adequate psychological theory must take this fact into account. How might functionalism do this? Straightforwardly. It must require of any state that is functionally equivalent to the sensation–of–warmth, say, that it have some instrinsic property or other whose presence is detectable by (= is causally sufficient for affecting) our mechanisms of introspective discrimination in such a way as to cause, in conceptually competent creatures, belief–states such as the belief that I have a sensation–of–warmth. If these sorts of causal relations are not part of a given state's functional identity, then it fails to be a sensation–of–warmth on purely functional grounds. (Sydney Shoemaker makes much the same point in Shoemaker, 1975. We do not know if he will agree with the points that follow.)

So functionalism *does* require that sensations have an intrinsic property that plays a certain causal role. But it is admittedly indifferent as to precisely what that intrinsic property might happen to be for any given type of sensation in any given person. So far as functionalism is concerned, that intrinsic property might be the spiking frequency of the signal in some neural pathway, the voltage across a polarized membrane, the temporary deficit of some neurochemical, or the binary configuration of a set of DC pulses. So long as it is one of these properties to which the mechanisms of introspective discrimination happen to be keyed, the property fills the bill.

"But *these* are not qualia!" chorus the outraged objectors. Are they not indeed. Recall the characterization of qualia given on the first page of this paper: " . . . those intrinsic or monadic properties of our sensations discriminated in introspection." Our sensations are anyway token–identical with the physical states that realize them, so there is no problem in construing a spiking frequency of 60 hertz as an intrinsic property of a certain sensation. And why should such a property, or any of the others listed, *not* be at the objective focus of introspective discrimination? To be sure, they would be *opaquely* discriminated, at least by creatures with a primitive self–conception like our own. That is to say, the spiking frequency of the impulses in a certain neural pathway need not prompt the non–inferential belief, "My pain has a spiking frequency of 60 Hz"; it may prompt only the belief, "My pain has a searing quality." But withal, the property you opaquely distinguish as "searingness" may be precisely the property of having 60 Hz as a spiking frequency.

There are many precedents for this sort of thing in the case of the intrinsic properties of material objects standardly discriminable in

observation. The redness of an object turns out to be a specific reflectance triplet for three critical wavelengths in the EM spectrum. The pitch of a singer's note turns out to be its frequency qua oscillation in air pressure. The warmth of a coffee cup turns out to be the vibrational energy of its molecules. The tartness of one's lemonade turns out to be its high relative concentration of H^+ ions. And so forth.

These chemical, electromagnetic, and micromechanical properties have been briskly discriminated by us for many millennia, but only opaquely. The reason is that we have not possessed the concepts necessary to make more penetrating judgments, and our mechanisms of sensory discrimination are of insufficient resolution to reveal on their own the intricacies uncovered by other means. Unambiguous perception of molecular KE, for example, would require sensory apparatus capable of resolving down to about 10^{-10} metres, and of tracking particles having the velocity of rifle bullets, millions of them, simultaneously. Our sensory apparatus for detecting and measuring molecular KE is rather more humble, but even so it connects us reliably with the parameter at issue. Mean molecular kinetic energy may not seem like an observable property of material objects, but most assuredly it is. (For a working–out of these themes in detail, see Paul Churchland, 1979.)

Similarly, spiking frequency may not seem like an introspectible property of sensations, but there is no reason why it should not be, and no reason why the epistemological story for the faculty of inner sense should be significantly different from the story told for outer sense. Qualia, therefore, are not an ineffable mystery, any more than colors or temperatures are. They are physical features of our psychological states, and we may expect qualia of some sort or other in any physical system that is sufficiently complex to be functionally isomorphic with our own psychology. The qualia of such a robot's states are not "absent." They are merely *unrecognized* by us under their physical/electronic description, or as discriminated by the modalities of outer rather than inner sense.

We may summarize all of this by saying that the functionalist need not and perhaps should not attempt to deny the existence of qualia. Rather, he should be a realist about qualia—in particular, he should be a *scientific* realist.

It is important to appreciate that one can be reductionistic about qualia, as outlined above, without being the least bit reductionistic about the taxonomy of states appropriate to psychological theory.

Once qualia have been denied a role in the type–identity of psychological states, the path described is open. If this line on qualia is correct, then it vindicates Ned Block's prophecy (1978: p. 309) that the explication of the nature of qualia does not reside in the domain of psychology. On the view argued here, the nature of specific qualia will be revealed by neurophysiology, neurochemistry, and neurophysics.

C. The Problem of Distinguishing States With Qualia From States Without

One could distinguish many differences between the sensations and the propositional attitudes, but one particular difference is of special interest here. A sensation–of–warmth, for example, has a distinct qualitative character, whereas the belief–that–Tom–is–tall does not. Can functionalism account for the difference?

Yes it can. The picture to be avoided here depicts sensations as dabbed with metaphysical paint, while beliefs remain undabbed and colorless. The real difference, we suggest, lies less in the objective nature of sensations and beliefs themselves, than in the nature of the introspective mechanisms used to discriminate and identify the states of each class. This hypothesis requires explanation.

How many different types of sensation are there? One hundred? One thousand? Ten thousand? It is difficult to make an estimate, since most sensations are arrayed on a qualitative continuum of some sort, and it is to some extent arbitrary where and how finely the lines between different kinds are drawn. It is plain, however, that the number of distinct continua we recognize, and the number of significant distinctions we draw within each, is sufficiently small that the brain can use the following strategy for making non–inferential identifications of sensations.

Consider the various physical properties which in you are characteristic of the repeatable brain state that realizes a given sensation. Simply exploit whichever of those physical properties is accessible to your innate discriminatory mechanisms, and contrive a standard habit of conceptual response ("lo, a sensation of warmth") to the property–evoked activation of those mechanisms. While this strategy will work nicely for the relatively small class of sensations, it will not work at all well for the class of beliefs, or for any of the other propositional attitudes. The reason is not that the brain state that realizes a certain belief *lacks* intrinsic properties characteristic of it alone. Rather, the reason is that there are far too many beliefs,

actual and possible, for us to have any hope of being able to discriminate and identify all of them on such a one–by–one basis. The number of possible beliefs is at least a denumerable infinity, and the number of possible beliefs expressible in an English sentence of ten words or less is still stupendous. Assuming a vocabulary of 10^5 words for English, the number of distinct strings of ten words or less is 10^{50}. Assuming conservatively that only one string in every trillion trillion is grammatically and semantically well–formed, this still leaves us over 10^{25} distinct sentences. Even if there were a distinct and accessible monadic property for each distinct belief–state, therefore, the capacity of memory is insufficient to file all of them. Evidently the brain must use some more systematic strategy for discriminating and identifying beliefs—a strategy that exploits in some way any belief's unique combinatorial structure.

But this is a very complex and sophisticated matter, requiring the resources of our higher cognitive capacities, capacities tuned to the complex relational, structural, and combinatorial features of the domain in which the discriminations are made. Unlike the sensation case, no narrow range of stimulus/response connections will begin to characterize the mechanisms at work here.

Sensations and beliefs, accordingly, must be introspectively discriminated by entirely distinct cognitive mechanisms, mechanisms facing quite different problems and using different strategies for their solutions. Sensations are identified by way of their intrinsic properties; beliefs are identified by way of their highly abstract structural features. It should not be wondered at then, that there is a subjective contrast in the nature of our awareness of each.

D. The Differentiation Problem

This problem arises because we are occasionally able to discriminate between qualitatively distinct sensations where we are ignorant of any corresponding functional differences between them, and even where we are wholly ignorant of the causal properties of both of them, as when they are new to us, for example. These cases are thought to constitute a problem in that functional considerations should bid us count the states as type–identical, whereas by hypothesis, they are type–distinct (see Block, 1978: p. 300).

The objection has two defects. First, sheer ignorance of functional differences need not bind us to counting the sensations as functionally identical. The functionalist can and should be a realist about functional properties. Functional identities are not

determined by what we know or do not know, but by what is actually out there in the world (or, *in* there, in the world). Second, the objection begs the question against functionalism by assuming that a discriminable qualitative difference between two sensations entails that they are type–distinct qua psychological states. We have already seen that this inference is wrong in any case: pains display a variety of qualitative characters, but because of their functional similarities, they still count as pains.

In short, we can and do make discriminations among our sensations in advance of functional understanding. But whether the discriminations thus made mark a difference of any importance for the taxonomy of psychological theory is another question. In some cases they will; in other cases they will not. What decides the matter is whether those qualitative differences mark any causal or functional differences relevant to the explanation of psychological activity and overt behavior.

So long as introspectible qualia were thought to be ineffable, or epiphenomenal, or dualistic, or essential for type–identity, one can understand the functionalist's reluctance to have anything to do with them. But once we have seen how the functionalist can acknowledge them and their epistemic role, within a naturalistic framework, the reluctance should disappear. For the taxonomy of states appropriate for psychological theory remains dictated entirely by causal and explanatory factors. Qualia are just accidental hooks of opportunity for the introspective discrimination of *dynamically* significant states.

II. *The Problem of Non–Standard Realizations*

Some of the issues arising here have already been broached in the section concerning absent qualia. However, novel problems arise as well, and organization is best served by a separate section. All of the problems here begin with the functionalist's central contention that the functional organization necessary and sufficient for personhood is an abstract one, an organization realizable in principle in an indefinite variety of physical systems. Such liberalism seems innocent enough when we contemplate the prospect of humanoid aliens, biomechanical androids, and electromechanical robots whose physical constitutions are at least rough parallels of our own. Who could deny that C3PO and R2D2 are persons? But our liberal intuitions are quickly flummoxed when we consider bizarre physical systems which might nevertheless realize the abstract causal

organization at issue, and such cases move one to reconsider one's generosity in the more familiar cases as well.

The following discussion will explore but two of these non–standard "persons": Ned Block's "Chinese Nation" (Block, 1978), and John Searle's "Chinese–speaking room" with the monolingual anglophone locked inside it (Searle, 1980). Block is concerned with the absence of *qualia* from states posing as sensations, and Searle is concerned with the absence of *intentionality* from states posing as propositional attitudes.

Block's example will be examined first. He has us imagine a certain Turing machine T_m, which is realized in the population of China, as follows. Each citizen enjoys a two–way radio link to a certain robotic device with sensory transducers and motor effectors. This robot is the body of the simulated person, and it interacts with its collective brain thus: it sends a sensory input message I_j to every citizen and subsequently receives a motor output message O_i from exactly one citizen. Which citizen sends what output is determined as follows. Overhead from a satellite some state letter S_k is displayed in lights for all to see. For each possible state letter S_k there is assigned a distinct subset of the population. In the rare event when S_k is displayed *and* input I_j is received, one person in the S_k group to whom I_j has been assigned performs this pre–assigned task: she sends to the robot the unique output message O_i antecedently assigned to her, radios the satellite to display the state letter S_p antecedently assigned to her, and then subsides, waiting for the next opportunity to do exactly the same thing in exactly the same circumstances.

As organized above, each citizen realizes exactly one square of the machine table that specifies T_m. (A machine table is a matrix or checkerboard with state letters heading the columns and input letters heading the rows. Any square is the intersection of some S_k and I_j, and it specifies an output O_i and a shift to some state S_p, where possibly $p = k$.) Block asks us to assume that T_m adequately simulates your own functional organization. One is likely to grant him this, since any input–output function can in principle be simulated by a suitable Turing machine. In pondering an apparently fussy detail, Block wonders, "How many homunculi are required?" and answers, "Perhaps a billion are enough; after all, there are only about a billion neurons in the brain" (p. 278). Hence his choice of China as the potential artificial brain.

Finally, Block finds it starkly implausible to suppose that this realization of T_m has states with a qualitative character like pains, tastes, and so on. It is difficult not to agree with him. His homunculi do not even interact with one another, save indirectly through the

satellite state letter, and, even less directly, through the adventures of the robot body itself. The shimmering intricacies of one's inner life are not to be found here.

The way to avoid this criticism is just to insist that any subject of beliefs and sensations must not only be "Turing–equivalent" to us (that is, produce identical outputs given identical inputs), it must be computationally equivalent to us as well. That is, it must have a system of inner states whose causal interconnections mirror those in our own case. This is not an arbitrary restriction. Folk psychology is, and scientific psychology should be, realistic about our mental states, and mere parity of gross behavior does not guarantee parity of causal organization among the states that produce it. The computational organization displayed in the Chinese Turing machine is not even distantly analogous to our own. If it were analogous to our own, worries about absent qualia could be handled as outlined in section I, part B, above.

There is a further reason why it is not arbitrary to demand a computational organization more along the lines of our own, and we may illustrate it by examining a further defect in Block's example. It is demonstrable that no T_m realized as described in the population of China could possibly simulate your input–output relations. There are not nearly enough Chinese—not *remotely* enough. In fact, a spherical volume of space centered on the Sun and ending at Pluto's orbit packed solidly with cheek–to–cheek Chinese (roughly 10^{36} homunculi) would still not be remotely enough, as we shall show.

Being realized on a one–man/one–square basis, the Chinese T_m can have at most 10^9 distinct possible outputs, and at most $10^9/S$ distinct possible inputs, where S is the total number of distinct state letters. That is, T_m has rather less than 10^9 possible inputs.

How many distinct possible inputs characterize your own functional organization? Since the present argument requires only a lower limit, let us consider just one of your retinas. The surface of your retina contains roughly 10^8 light sensitive cells, which we shall assume (conservatively) to be capable of only two states: stimulated or unstimulated. Good eyesight has a resolution limit of about one foot at a distance of a mile, or slightly less than one arc–minute, and this angle struck out from the lens of the eye subtends about six microns at the retina. This is roughly the distance between the individual cells to be found there, so it is evident that individual cells, and not just groups, can serve as discriminative atoms, functionally speaking.

If we take distinct stimulation patterns in the set of retinal cells as distinct inputs to the brain, it is evident that we are here dealing with

2 to the (10^8) power distinct inputs. This is an appallingly large number. Since $2^{332} \approx 10^{100}$ (a googol), $2^{10^8} \approx 10^{30,000,000}$ distinct possible inputs from a single retina! Since a one–man/one–square Turing machine must have at least as many homunculi as possible inputs, any such realization adequate to the inputs from a single retina would require no less than $10^{30,000,000}$ distinct homunculi. However, there are only about 10^{80} distinct atoms in the accessible universe. Small wonder the Chinese nation makes an unconvincing simulation of our inner lives, but we should never have acquiesced in the premise that a Turing machine thus realized could even begin to simulate your overall functional organization. The Chinese robot body can have at most a mere 30 binary input sensors, since $2^{30} \approx 10^9$, and the number of inputs cannot exceed 10^9.

This argument does not depend on inflated estimates concerning the retina or its input to the brain. (It might be objected that retinal cell stimulation is not independent of the state of its immediate neighbor cells, or that the optic nerve has only 800,000 axons.) If your retina contained only 332 discriminatory units, instead of 10^8, the number of distinct inputs would still be 2^{332}, or roughly 10^{100}: ninety–one orders of magnitude beyond the capacity of the Chinese nation, and twenty orders of magnitude beyond the atoms in the universe. Nor have we even begun to consider the other dimension of the required machine table: the range of states, S, of the brain which receives these inputs, a brain which has at least 10^{10} distinct cells in its own right, each with about 10^3 connections with other cells. Our estimate of the number of distinct states of the brain must be substantially in excess of $10^{30,000,000}$, our number for the retina.

Our conclusion is that *no* brute–force one–device/one–square realization of a Turing machine constructible in this universe could even begin to simulate your input–output organization. Even the humblest of creatures are beyond such simulation. An unprepossessing gastropod like the sea slug *Aplysia Californica* has well in excess of 332 distinct sensory cells, and thus is clearly beyond the reach of the crude methods at issue. This does not mean that the human input/output relations cannot be represented by an abstract Turing machine T_m. What it does mean is that any *physical* machine adequate to such simulation *must* have its computational architecture and executive hardware organized along lines vastly different from, and much more unified and efficient than, those displayed in Block's example. That example, therefore, is not even remotely close to being a fair test of our intuitions. Quite aside from the question of qualia, the Chinese Turing machine couldn't simulate an earthworm.

This weakness in the example is not adequately made up by allowing, as Block does at one point (p. 284), that each homunculus might be responsible for a wide range of inputs, each with corresponding outputs. On this modification, each homunculus would thus realize, dispositionally, many machine–table squares simultaneously. Suppose then that we make each Chinese citizen responsible for one billion squares peculiar to him. This raises the number of distinct inputs processible by the system to 10^9 citizens $\times$ 10^9 squares $= 10^{18}$ possible inputs, still well short of the $10^{30,000,000}$ we are striving for. Well, how many squares must each citizen realize if the nation as a whole *is* to instantiate some Turing machine adequate to handle the required input? The answer is, of course, $10^{(30,000,000-9)}$ squares each. But how will each citizen/homunculus handle this awesome load? *Not* by being a simple one-device/one-square Turing machine in turn, as we have already seen. No physical simulation adequate to your input–output relations, therefore, can avoid having the more unified and efficient modes of computational organization alluded to in the last paragraph, even if they only show up as modes of organization of its various subunits. Any successful simulation of you, that is, must somewhere display a computational/executive organization that is a much more plausible home for qualitative states than Block's example would suggest.

But can a number of distinct persons or near–persons collectively constitute a further person? Apparently so, since the system consisting of your right hemisphere and left hemisphere (and your cerebellum and thalamus and limbic system, etc.) seems to do precisely that. Further attempts to construct homunculi–headed counter–examples to functionalism should perhaps bear this fact in mind.

The argument of the preceding pages does not, of course, show that the specific details of *our* computational organization are essential to achieving the informational capacity required. And this raises a question we might have asked anyway: if we do require of any subject of sensations, beliefs, and so forth, that it be functionally equivalent to us in the strong sense of "computationally equivalent," do we not then run the opposite danger of allowing too *few* things to count as sites of mentality? (see again Block, p. 310ff.) If we restrict the application of the term 'mentality' to creatures having sensations, beliefs, intentions, etc., we will indeed have become too restrictive. Yet the functionalist need not pretend that our internal functional organization exhausts the possible kinds of mentality. He

need only claim that our kind of internal functional organization is what constitutes a psychology of *beliefs*, *desires*, *sensations*, and so forth. He is free to suggest that an alien functional organization, comparable only in sophistication to our own, could constitute an alien psychology of quite different internal states. We could then speak of Martian mentality, for example, as well as of human mentality.

Still, it might be wondered, what is the shared essence that makes both of us instances of the now more general term, 'mentality'? There need be none, beyond the general idea of a sophisticated control center for complex behavior. One of the functionalist's principal theses, after all, is that there are no natural essences to be found in this domain. If he is right, it is folly to seek them. In any case, it is question–begging to demand that he find them.

On the other hand, there may yet prove to be some interesting natural kind, of which both we and the Martians are variant instances: some highly abstract thermodynamic kind, perhaps. In that case, orthodox functionalism would be mistaken in one of its purely negative theses. On this matter, see the final section of this paper.

Let us now turn to John Searle's worries about meaning and intentionality. The states at issue here are beliefs, thoughts, desires, and the rest of the propositional attitudes. On the functionalist's view, the type–identity of any of these states is determined by the network of relations it bears to the other states, and to external circumstances and behavior. In the case of the propositional attitudes, those relations characteristically reflect a variety of logical and computational relations among the propositions they "contain." We can thus at least imagine a computer of sufficient capacity programmed so as to display an economy of internal states whose interconnections mirror those in our own case. The simulation would create the required relational order by exploiting the logical and computational relations defined over the formal/structural/combinatorial features of the individual propositional states.

Searle has no doubt that such a simulation could, in principle, be constructed. His objection to functionalism is that the states of such a system would nevertheless lack meaning and intentionality: " . . . no purely formal model will ever be sufficient by itself for intentionality because the formal properties are not by themselves constitutive of intentionality" (Searle, 1980: p. 422). His reasons for holding this position are illustrated in the following thought–experiment.

Imagine a monolingual anglophone locked in a room with (a) a substantial store of sequences of Chinese symbols, and (b) a set of complex transformation rules, written in English, for performing operations on sets and sequences of Chinese symbols. The occupant periodically receives a new sequence of Chinese symbols through a postal slot. He applies his transformation rules dutifully to the ordered pair, ⟨new sequence, old store of sequences⟩ , and they tell him to write out a further sequence of Chinese symbols, which he sends back out through the postal slot.

Now, unknown to the occupant, the large store of sequences embodies a rich store of information on some one or more topics, all written in Chinese. The new sequences sent through the door are questions and comments on those topics. The transformation rules are a cunningly devised program designed to simulate the thought processes and conversational behavior of a native speaker of Chinese. The symbol–sequences the occupant sends out are "responses" to the queries and comments received. We are to suppose that the transformation rules are well–devised, and that the simulation is as convincing as you please, considered from the outside.

However convincing it is, says Searle, it remains plain that the room's occupant does not understand Chinese: he applies transformation rules, and he understands those rules, but the sequences of Chinese symbols are meaningless to him. Equally clear, claims Searle, is that the system of the room–and–its–contents does not understand Chinese either. Nothing here understands Chinese, save those sending and receiving the messages, and those who wrote the program. No computational state or output of that system has any meaning or intentionality save as it is interpretively imposed from without by those who interact with it.

However, concludes Searle, this system already contains everything relevant to be found in the physical realization of any purely formal program. If meaning and intentionality are missing here, they will be missing in any such attempt to simulate human mental activity. Instantiating a program could not be a sufficient condition of understanding,

> Because the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even *symbol* manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is

> solely in the minds of those who program them, those who send in the input and those who interpret the output. [Ibid., p. 422]

The set of commentaries published in the same issue provides many useful and interesting criticisms of Searle's argument, and of his conclusion as well. The critical consensus is roughly as follows. If the system of the room–plus–contents is upgraded so that its conversational skills extend beyond a handful of topics to include the entire range of topics a normal human could be expected to know; and if the system were supplied with the same inductive capacities we enjoy; and if the "belief store" were integrated in the normal fashion with some appropriately complex goal structure; and if the room were causally connected to a body in such a fashion that its inputs reflected appropriate sensory discriminations and its outputs produced appropriate behavior; then the system of the room–plus–contents jolly well would understand Chinese, and its various computational states—beliefs that p, desires that q—would indeed have meaning and intentionality, in the same way as with a normal Chinese speaker.

Searle is quite willing to consider upgradings of the kind described—he attempts to anticipate them in his paper—but he is convinced they change nothing relevant to his case. As it emerges clearly in his "Author's Response" (pp. 450–56), of central importance to his argument is the distinction between

> . . . cases of what I will call *intrinsic intentionality*, which are cases of actual mental states, and what I will call *observer–relative ascriptions* of intentionality, which are ways that people have of speaking about entities figuring in our activities but lacking intrinsic intentionality. [p. 451] [The latter] are always dependent on the intrinsic intentionality of the observers. [p. 452]

Examples of the latter would be the words and sentences of one's native tongue. These have meaning and intentionality, allows Searle, but only insofar as they bear certain relations to our beliefs, thoughts, and intentions—states with intrinsic intentionality. A simulation of human mentality grounded in a formal program may yield states having this derivative observer–relative brand of intentionality, concedes Searle, but they cannot have intrinsic intentionality. And since they lack a feature essential to genuine mental states, they cannot be genuine mental states, and to that extent the simulation must be a failure.

As we see it, this criticism of functionalism is profoundly in error. It is a mistake to try to meet it, however, by continuing with the

strategy of trying to upgrade the imagined simulation in hopes of finally winning Searle's concession that at last its states have achieved intrinsic intentionality. The correct strategy is to argue that our own mental states are just as innocent of "intrinsic intentionality" as are the states of any machine simulation. On our view, all ascriptions of meaning or propositional content are relative—in senses to be explained. The notion of "intrinsic intentionality" makes no more empirical sense than does the notion of position in absolute space. We shall try to explain these claims.

There are basically just two ways in which one can assign propositional content to the representational states of another organism. An example of the first is the translation of a foreign language. An example of the second is the calibration of an instrument of measurement or detection.

In the case of translation, one assigns specific propositional contents to the alien representations because one has found a general mapping between the alien representations and our own such that the network of formal and material inferences holding among the alien representations closely mirrors the same network holding among our own representations. Briefly, their collected representations display an *intensional structure* that matches the intensional structure displayed by our own.

The story is essentially the same when we are assigning propositional content to an alien's thoughts, beliefs, etc. It matters naught whether the alien's representation is overt, as with a sentence, or covert, as with a belief. We assign a specific content, *p*, to one of the alien's representations, on the strength of whatever assurances we have that his representation plays the same abstract inferential role in his intellectual (computational) economy that the belief–that–*p* plays in ours. And what goes for aliens goes also for one's brothers and sisters.

This is not to say that the representational states of other humans have content only insofar as others interpret them in some way. After all, the set of abstract inferential relations holding among the representations in someone's intellectual economy is an objective, non–relational feature of that person. But it does mean that the content, call it the *translational content*, of any specific representation of his *is a matter of the inferential/computational relations it bears to all the rest of his representations*. There can be no question of an isolated state or token possessing an intrinsic translational content; it will have a specific translational content only if, and only insofar as, it enjoys a specific set of relations to the other elements in a *system* of representations.

Contrast translational content with what is naturally termed *calibrational content*. The repeatable states of certain physical systems are more–or–less reliable indicators of certain features of their environment, and we may assign content (e.g., 'The temperature is O°C'.) to such states (e.g., a certain height in a column of red alcohol) on the strength of such empirical connections. This goes for the human system as well. The various states we call 'perceptual beliefs' can be assigned contents in this manner, as a function of which type of environmental circumstance standardly triggers their occurrence. In fact, if a system has any systematic responses to its environment at all, then calibration can take place even where translation cannot—either because the system simply lacks the internal economy necessary for translational content, or because the intensional structure of that economy is incommensurable with our own. Furthermore, calibrational content may regularly *diverge* from translational content, even where translation is possible. Consider an utterance which calibrates as 'There is thunder', but which translates as 'God is shouting'; or one which calibrates as 'This man has a bacterial infection', but which translates as 'This man is possessed by a pink demon'.

Accordingly, Searle is right to resist the suggestion that merely hooking up the room–system, via some sensors, to the outside world, would supply a unique meaning or content to the room's representational states. Genuine mental states do indeed have a content or intentionality that is independent of, and possibly quite different from, their calibrational content. (The reader will notice that this entails that it is just possible that all or most of our beliefs are false—that their translational contents may be systematically out of agreement with their actual calibrational contents. For an extended exploration of this real possibility, see Paul Churchland, 1979: ch. 2. On this matter see also Stephen Stich, this issue.) That independent intentionality is their *translational content*. But this content falls well short of being the "intrinsic intentionality" Searle imagines our states to have. Translational content is not environmentally determined, nor is it observer–relative, but it is most certainly a *relational* matter, a matter of the state's inferential/computational relations within a system of other states. Accordingly, it is entirely possible for translational content to be possessed by the states of a machine—the realization of a purely formal program.

What more than this Searle imagines as fixing the content of our mental states, we are unable to surmise. He floats the distinction

between intrinsic intentionality and other kinds by means of illustrative examples (p. 451); he hazards no palpable account of what intrinsic intentionality consists in; and the intuitions to which he appeals can be explained in less mysterious ways, as outlined above. To conclude, there is simply no such thing as intrinsic intentionality, at least as Searle conceives it. Functionalists need not be concerned then that computer simulations of human mentality fail to display it.

We complete this section by underscoring a contrast. In the first half of this section we conceded to the critic of functionalism that our mental states have qualia, but we argued that the states of a machine simulation could have them as well. In this second half we have conceded to the critic of functionalism that the states of certain machine simulations must lack intrinsic intentionality. But we insist that our own states are devoid of it as well.

III. *Functionalism and Methodology*

Despite the defenses offered above, we do wish to direct certain criticisms against functionalism. The criticisms are mainly methodological rather than substantive, however, and we shall here provide only a brief summary, since they have been explained at length elsewhere.

A. Conceptual Conservatism

No functionalist will suppose that the functional organization recognized in the collected lore of folk psychology exhausts the functional intricacies that make up our internal economy. All will agree that folk psychology represents only a partial, and in some respects even a superficial, grasp of the more complex organization that empirical psychology will eventually unravel. Even so, there is a decided tendency on all sides to suppose that, so far as it goes, folk psychology is essentially correct in the picture it paints, at least in basic outlines. Empirical psychology will add to it, and explain its principles, most expect, but almost no one expects it to be overthrown or transmogrified by such research.

This sanguine outlook is not unique to functionalists, but they are especially vulnerable to it. Since the type–identity of mental states is held to reside, not in any shared physical or other natural essences, but in the structure of their causal relations, there is a tendency to construe the generalizations connecting them as collectively *stipulating* what it is to be a belief, a desire, a pain, and so forth. Folk psychology is thus removed from the arena of empirical criticism.

This is unfortunate, since the "denaturing" of folk psychology does not change its epistemic status as a speculative account of our internal workings. Like any other theory, it may be radically false; and like any other deeply entrenched theory, its falsity is unlikely to be revealed without a vigorous exploration of that possibility.

A functionalist can of course accept these points without danger to what is basic in his position. Nevertheless, they are worth making for two reasons. First, eliminative materialism is not a very widespread opinion among adherents of functionalism, despite being entirely consistent with their view. And second, there are very good reasons for doubting the integrity of folk psychology, in its central structures as well as in its peripheral details (see Paul Churchland, 1979: ch. 5; 1981a; also, Patricia Churchland, 1980a; 1980b; and look for Stephen Stich's *The Case Against Belief*, in progress).

B. Top–Down Versus Bottom–Up

Given that the essence of our psychological states resides in the set of causal relations they bear to one another, etc., and given that this abstract functional organization can be realized in a nomically heterogeneous variety of substrates, it is fair enough that the functionalist should be more interested in that abstract organization than in the machinery that realizes it. With the science of psychology, it is understanding the "program" that counts; an understanding of such hardwares as may execute it is secondary and inessential.

This much is fair enough, but so long as we are so profoundly ignorant of our functional organization as we are at present, and ignorant of where to draw the line between "hardware" and "program" in organisms, we cannot afford to be so casual about or indifferent to the neurosciences as the preceding rationale might suggest. If we wish to unravel the functional intricacies of our internal economy, one obvious way to go about it is to unravel the intricacies of the physical system that executes it. This "bottom–up" approach is not the only approach we might follow, but it does boast a number of advantages: it is very strongly empirical; it is not constrained by the preconceptions of folk psychology; it has the capacity to force surprises on us; it permits a non–behavioral comparison of cognitive differences across species; it enjoys direct connections with evolutionary ethology; and at least in principle it *can* reveal the functional organization we are looking for.

Neuroscience is an awkward and difficult pursuit, however, and

there is an overwhelming preference among philosophers, psychologists, and artificial intelligence researchers for a more "top–down" approach: hypothesize functional systems ("programs") and test them against our molar behavior, as conceived within common sense. This is entirely legitimate, but if the functionalist is moved by the argument from abstraction to ignore or devalue the bottom–up approach, his methodology is dangerously conservative and one–sided. (We have discussed these shortcomings at length in Patricia Churchland, 1980a, 1980c; and in Paul Churchland, 1981a, 1981b, 1980.)

C. Reductionism

Thanks to the argument from abstraction, functionalists tend to be strongly anti–reductionist. They deny that there can be any general characterization of what makes something a *thinker* that is expressible in the language of any of the physical sciences. Given the variety of possible substrates (biological, chemical, electromechanical) that could realize a thinking system, it is difficult not to agree with them. But it does not follow, from multiple instantiability *per se*, that no such general characterization is possible. It follows only that the required characterization cannot be expressed in the theoretical vocabulary peculiar to any one of the available substrates. It remains entirely possible that there is a level of physical description sufficiently abstract to encompass all of them, and yet sufficiently powerful to fund the characterization required.

As it happens, there is indeed a physical theory of sufficient generality to encompass the activity of all of these substrates, and any others one might think of. The theory is thermodynamics—the general theory of energy and entropy. It has already supplied us with a profoundly illuminating characterization of what the nineteenth century called "vital activity," that is, of the phenomenon of life. And it is far from unthinkable that it might do the same for what this century calls "mental activity." (For a brief exploration of these ideas, see Paul Churchland, 1981b.) The theoretical articulation of such a characterization would be a very great achievement. It would be unfortunate if the search for it were impeded by the general conviction that it is impossible, a conviction born of the anti–reductionist urgings of a false orthodoxy among functionalists.

REFERENCES

Block, Ned. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science*, Vol. IX: pp. 261–325. Edited by C. Wade Savage. Minneapolis: University of Minnesota Press, 1978.

Block, Ned and Fodor, Jerry. "What Mental States Are Not." *The Philosophical Review*, LXXXI (1972): 159–81.

Churchland, Patricia Smith. (1980a). "A Perspective on Mind–Brain Research." *The Journal of Philosophy*, LXXVII, 4 (April, 1980): 185–207.

——— (1980b). "Language, Thought, and Information Processing." *Noûs* 14 (1980): 147–69.

——— (1980c). "Neuroscience and Psychology: Should the Labor Be Divided?" *The Behavioral and Brain Sciences*, III, 1 (March, 1980): 133.

Churchland, Paul M. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge: Cambridge University Press, 1979.

——— (1980). "Plasticity: Conceptual and Neuronal." *The Behavioral and Brain Sciences*, III, 1 (March, 1980): 133–34.

——— (1981a). "Eliminative Materialism and the Propositional Attitudes." *The Journal of Philosophy* (LXXVIII, 2 February, 1981): 67–90.

——— (1981b). "Is *Thinker* a Natural Kind?" *Dialogue* (forthcoming, 1981).

Putnam, Hilary. "The Nature of Mental States." *Materialism and the Mind–Body Problem*, pp. 150–61. Edited by David Rosenthal. New Jersey: Prentice–Hall, 1971.

Searle, John. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences*, III, 3 (September, 1980): 417–57.

Sellars, Wilfrid. "Empiricism and the Philosophy of Mind." *Minnesota Studies in the Philosophy of Science*, Vol. I. Edited by Herbert Feigl and Michael Scriven. Minneapolis: University of Minnesota Press, 1956. Reprinted in Sellars, *Science, Perception, and Reality*. London: Routledge & Keegan Paul, 1963: 127–96.

Shoemaker, Sydney. "Functionalism and Qualia." *Philosophical Studies* 27 (1975): 291–315.

Stich, Stephen. *The Case Against Belief*. In progress.