

WHAT IS IT LIKE TO BE A HUMAN (INSTEAD OF A BAT)?

My purpose in this paper is to discuss and defend an objection to physicalist or materialist accounts of the mind—one that I believe to be essentially conclusive.^[1] The argument in question is not new. A version of it seems to be lurking, along with much else, in Thomas Nagel's famous paper "What Is It Like to Be a Bat?"^[2]; and a somewhat more explicit version is to be found in a well-known paper by Frank Jackson.^[3] Despite the efforts of Nagel and Jackson (and some others), however, I believe that the most compelling version of the argument has not emerged clearly, with the result that responses that in fact fail to speak to its central point are widely taken to be adequate. Thus one purpose of the present paper is to offer what I regard as a more perspicuous restatement of the Nagel-Jackson argument, one which shows clearly why the responses in question do not work. A second purpose is to suggest that the application of the argument is in fact very much wider than the case of phenomenal properties or qualia upon which both Nagel and Jackson focus, that it in fact applies just as well to the content of intentional mental states like thoughts and indeed to the general phenomenon of consciousness itself.

I

I begin with a brief and selective recapitulation of Nagel's and Jackson's presentations of the argument and of some of the critical responses they evoked, focusing on those raised by Paul Churchland.

Though, as we will see, there are several other balls in the air, the strand of Nagel's argument upon which I wish to focus goes at least approximately as follows: It is reasonable to assume that bats have conscious experience of some kind, that as Nagel puts it "there is something it is like to be a bat" [423]. But such experience is surely enormously different from our own in many ways, due to the very different "range of activity and sensory

apparatus" [423] possessed by bats, in particular their well-known capacity of perceiving the world and navigating through it via a kind of sonar resulting from the reflection of their own high-pitched cries. Nagel's question is whether we could ever come to know "what it is like to be a bat," and in particular whether we could do this on the basis of a thoroughgoing knowledge of the physical or material facts pertaining to bat physiology. His claim is that we could not, and the suggested conclusion, which he himself never quite draws, is that physicalism is false.

As will emerge, I believe that the foregoing argument is essentially sound. I also believe that it is present in Nagel. But it is very hard to be sure of the latter claim, because so many other ideas and suggestions are present in the paper as well, ideas and suggestions that seem in some cases to be incompatible with the foregoing argument and in other cases to point in at least rather different directions. There is, first, the idea of a "point of view," with the suggestion that certain kinds of facts may be knowable only from a certain point of view and the accompanying distinction, suggestive but also quite elusive, between various "subjective" points of view and the "objective" point of view characteristic of physical science. Secondly, there is the concern with conceptual limitations, and the suggestion that the main problem with regard to bats is that we may not have and may not be able to acquire the right concepts to capture bat experiences. Third, there is the suggestion that the right conclusion is not so much that physicalism is false as that we do not understand how it could be true—which might still be compatible, Nagel suggests, with having good or even compelling reasons to think that it is true. And, fourth, even the formulation that I have echoed in my title in terms of "what it is like to be a bat" is at least potentially misleading, in that it (along with the employment of the objective/subjective dichotomy) might suggest that the knowledge that we are lacking with respect to bats is not so much knowledge of *facts* as knowledge of what it would "feel like from the inside" to be a bat—thereby inviting the

reasonably plausible response that this is not a sort of knowledge that physicalism could be expected to provide even if it were true.[\[4\]](#)

I do not mean to suggest that some or all of these ideas may not be valuable on their own. In particular, the idea of subjective and objective points of view may well yield a much deeper insight into the mysterious nature of consciousness than could ever be derived from the argument that I mean to focus on here. My point is merely that these elements are quite inessential to the argument that is my present concern, an argument that I believe to be more immediately and unproblematically compelling as an objection to physicalism or materialism, even if perhaps ultimately less insightful in other ways. By diverting attention, those other elements tend to prevent that argument from emerging clearly and also to invite irrelevant responses.

Jackson's version of the argument focuses more clearly on the central point. In the most compelling version, it imagines a brilliant neurophysiologist, Mary, who lives her entire life, acquires her education, and does all of her scientific work in a black-and-white environment, using black-and-white books and black-and-white television for all of her learning and research. In this way, we may suppose, she comes to have a complete knowledge of all the physical facts in neurophysiology and related fields, together with their deductive consequences, insofar as these are relevant—thus arriving at as complete an understanding of human functioning as those sciences can provide. In particular, Mary knows the functional roles of all of the various neurophysiological states, including those pertaining to sense perception, insofar as these are reflected in their causal relations to sensory inputs, behavioral outputs, and other such states. But despite all of this knowledge, Jackson suggests, Mary does not know all that there is to know about human mental states: for when she is released from her black-and-white environment and allowed to view the world normally, she will, by viewing objects like ripe tomatoes, "learn what it is like to see something red,"[\[5\]](#) and analogous things about other colors. "But then," comments Jackson,

"it is inescapable that her previous knowledge was incomplete. But she had all the physical information. Ergo there is more to have than that, and Physicalism is false."[\[6\]](#)

While this version of the argument is certainly less burdened with other distractions than Nagel's and is also once again, I believe, essentially sound, there are still problems. These may be examined by considering two objections to the Jackson version offered by Churchland.[\[7\]](#)

Churchland's first objection is that while Mary undoubtedly learns *something* new when she is released from her black-and-white environment, what she acquires is not *knowledge* in the same sense of that term in which it is a consequence of the truth of physicalism that she already knows all there is to know. The sense of "knowledge" in which physicalism guarantees that Mary's knowledge is complete is "a matter of having mastered a set of sentences or propositions," while the sort that Mary acquires is:

a matter of having a representation of redness in some prelinguistic or sublinguistic medium of representation for sensory variables, or . . . a matter of being able to make certain sensory discriminations, or something along these lines. [\[23\]](#)

Churchland's point may perhaps be put somewhat more clearly by saying that it is not the case, according to him, that what Mary is initially lacking and then later comes to acquire is a knowledge of certain *facts* or *truths* about human mental states, but that knowledge of facts or truths is all that the physical account could be expected to supply, even if physicalism were correct. I believe that Churchland's claim that Mary does not come to know any new facts or truths is mistaken, but this point is difficult to establish clearly within the context of Jackson's formulation of the argument: if she learns new facts or truths, what exactly are they? Thus Jackson, in his response, is reduced to arguing in a very indirect fashion for the existence of such facts (by appealing to the genuineness of the problem of other minds).[\[8\]](#)

Churchland's second objection turns on the intriguing suggestion that Mary, once she has learned to employ the concepts of a completed neuroscience in introspection, might be able to imaginatively extrapolate from her introspective awareness of her black-and-white experiences to the experiences she would have if she were in the neurophysiological states corresponding to color experience and thus might come to know "what it is like to see something red." Jackson's response^[9] is that if physicalism were true, Mary should *know* what the experience is like, rather than merely having to imagine it. But I do not see why Churchland, if his point were otherwise sound, could not claim that Mary might indeed come to know in this way what the experience is like, albeit via a kind of imaginative inference rather than direct experience. Again, I believe that Churchland is mistaken here: both about what Mary would be able to do and, more importantly, about the relevance of this issue to the central point of the argument. But the way in which Jackson has formulated the argument makes it hard to clearly establish either of these things.

II

What is needed, in my judgment, is a version of the argument that (i) makes it clear that there are *facts* or *truths* about human mental states that someone in Mary's position does not and cannot know on the basis of purely physical and neurophysiological knowledge, however complete that may be, and (ii) avoids relatively intractable issues about what Mary might be able to imagine or imaginatively infer on the basis of her own experience. And the way to do this, I suggest, is in effect to invert Nagel's original example, in a way that he himself suggests in passing but does not develop [425]: instead of imagining ourselves trying to know or comprehend the experiences of an alien form of life, we need instead to imagine an alien form of life trying to know or comprehend our experiences.

Suppose then that a brilliant Martian scientist comes to earth to investigate, with our full cooperation, the nature and makeup of human beings. Being a Martian, he has, we

may suppose, a quite different sensory apparatus from ours, but one which is still quite adequate, given his complete mastery of the standard sorts of inductive and explanatory reasoning, to arrive at a complete knowledge of any purely physical phenomenon. Thus, in time, the Martian arrives at an ideally complete knowledge of the physical and neurophysiological facts concerning human beings, including those pertaining to causally defined functional roles. Does he thereby come to know all of the facts about human mental states such as experiences of color?

Suppose that I am one of the subjects studied by the Martian. On a particular occasion, I look at a newly mowed, well-watered, and healthy lawn and thereby have an experience of a certain specific phenomenal or sensuous property, one which is somewhere toward the middle of the range of such properties that I am accustomed to call "green." On another occasion, I look at a newly painted fire engine and thereby have an experience of a second specific phenomenal or sensuous property, one which is somewhere toward the middle of the range of properties that I am accustomed to call "red." It is, I submit, simply a fact about me in the most straightforward possible sense that on the first occasion I experience the first property and that on the second occasion I experience the second property. The Martian is present on both occasions and is carefully monitoring my physical and neurophysiological states with an elaborate set of instruments that he has devised for this purpose. He thereby comes to know everything about those states, including their causal relations to other states, to as fine a level of detail as could possibly be relevant.[\[10\]](#) Does he thereby know that I am experiencing the first property on the first occasion and the second property on the second occasion?

I have stipulated that the Martian does not possess senses like ours. In particular, he does not possess eyes and a faculty of vision like ours. Thus one thing that he cannot do is determine what property I am experiencing by looking at the relevant objects himself. Nor should he need to do this, since facts about his own experiences are of course no part of his

supposedly complete physical and neurophysiological account of humans in general and of me in particular. (The same thing is in fact true of Mary: Though she happens to be a member of the species that she is investigating, her introspective awareness of her own experiences is still not a part of the ideally complete physical and neurophysiological account of humans at which she arrives by the methods of physical science. This is why Churchland's speculations about her imaginative extrapolations are strictly irrelevant.)

The Martian does not experience colors in the way and in the contexts that we do. But it is still possible that he is familiar in some other way with the specific phenomenal or sensuous properties at issue, and it will help to focus the essential point if we suppose that this is so. Thus suppose that he does experience those very properties, albeit in some quite different causal context. Perhaps he experiences colors when he hears or otherwise senses vibrations in the air corresponding to music. Or, less fancifully, perhaps he does have something like eyes and vision, but in relation to a quite different range of electromagnetic radiation, and experiences all of the colors that we experience (and perhaps others?) in that connection. Thus, we may suppose, he has a perfectly good grasp of the *concepts* of having an experience of each of the two properties in question, and the issue is only whether he can apply those concepts correctly to me.[\[11\]](#)

We may even concede to the Martian one more useful piece of information, albeit one that he almost certainly could not in fact arrive at on his own. Let us stipulate not only that he is familiar with color properties and possesses the concepts of having such experiences, but even that he somehow knows[\[12\]](#)—perhaps God whispers it in his ear or appropriate alternative sense organ—that two specific color properties out of the ones with which he is familiar are in fact the two that I am experiencing on the two occasions in question (but not of course which is which). In addition, we may suppose that the Martian has solved the difficult but probably not entirely intractable problem of isolating the specific features of my neurophysiology that are relevant to the issue we are concerned with, so that he is able to

focus on two relatively restricted neurophysiological states that are, supposing that physicalism is true, identical to my experiencing of the two colors. Thus he is able to formulate to himself two pairs of propositions, one pair identifying the first of these restricted states with an experiencing of the first of the two properties and the second restricted state with an experiencing of the second of the two properties, and the other pair reversing these ascriptions. He thus knows, we are supposing, that the propositions in one pair are true and those in the other pair false, but not which is which. Can he tell, solely on the basis of his complete physical and neurophysiological knowledge, which is the correct pair?

In thinking about this question, it is important to be quite clear about the exact shape of the issue. If physicalism is true, I submit, then the Martian should not have to extrapolate or surmise or guess, in however educated a fashion, in order to determine which pair of propositions is the correct one. If the ideal physical and neurophysiological account is indeed a *complete* account of all the facts concerning humans and their mental states, and if one of the two pairs of propositions is true and the other false in relation to that subject-matter, then it seems to follow that the propositions of the true pair must be already included in some way in that account, and that the propositions in the other pair must be in some way incompatible with that account—where the inclusion and incompatibility in question can apparently be only *logical* or *analytic* inclusion or incompatibility. And this would apparently mean in turn that the ideas or concepts of the two phenomenal or sensuous properties in question would have to be either already present in the neurophysiological account or somehow strictly definable on the basis of neurophysiological ideas or concepts. The former of these alternatives seems clearly mistaken, which is just to say that neurophysiology does not explicitly invoke the idea of sensuous or phenomenal color. And the latter alternative is no more palatable. One way to argue this point is to appeal to the familiar view that color concepts are primitive or undefinable, a view that I believe to be

correct albeit somewhat elusive. But even apart from this sort of appeal, the idea that the concepts of the various sensuous or phenomenal colors are strictly definable on the basis of neurophysiological primitives has, if anything, even less plausibility than the old phenomenalist idea that physical object concepts are definable in purely sensory terms. I do not know how to strictly prove that no such definition is possible, but I know of no one who has ever seriously defended such a view, nor of any way to make it even minimally plausible.[\[13\]](#)

Thus it seems utterly plain that the answer to our original question is "no." All that the Martian's physical and neurophysiological knowledge can give him is increasingly complicated accounts of the structure of the two restricted neurophysiological states and of their structural and causal relations to each other and to other states and processes of the same kind. But all of this knowledge, however detailed and elaborate we may suppose it to be, would still be entirely compatible with the truth of either of the two pairs of propositions. The indicated conclusion is that although the Martian scientist knows all the physical and neurophysiological facts there are, he does not know all of the *facts* there are, and hence that physicalism or materialism is false.

I want to conclude this section by considering briefly two possible rejoinders on behalf of the physicalist, both of them attempts to evade the argument in the only way that might still seem open to him: via a denial that it is a consequence of the truth of physicalism that all of the facts about a given sort of thing must be logically contained in a complete physical account of that thing. It is obvious that many physicalists, at least since the death of logical behaviorism, have wanted to avoid such a requirement, indeed that this has been much of the point of the various physicalist positions. But it seems to me very doubtful that any adequate rationale for rejecting this requirement, as opposed to qualifying it in minor and ultimately irrelevant ways, has ever been given.

First. The most obvious rejoinder is that there are at least two conspicuous sorts of facts about a thing, one of them perhaps a subclass of the other, that need not, even if physicalism is true, be thus contained in a complete physical account that is confined to that thing. One sort of fact pertains to the *function* or *purpose* of the thing: thus I could know all of the purely physical facts about a certain sort of object and still not know that it is a chair, because being a chair has to do with its function for human beings and not with its purely physical description. The other sort of fact pertains to classifications that are relative to human needs or purposes and perhaps also to some degree conventional or even arbitrary: thus, e.g., I could know all of the physical facts about a thing, including the precise mean kinetic energy of its molecules, and still not know that it was hot rather than lukewarm, as classified by common sense, because the difference here has to do with a fuzzy and relatively arbitrary line that humans draw, for reasons having to do with their own bodily temperature, within an essentially continuous range of physical temperatures.

But what makes facts of these kinds (and perhaps others of similar sorts as well) unknowable on the basis of a complete physical description of the thing is that they implicitly have to do with *relations* between the thing in question and other things, in this case humans and their purposes and classifications, and it is obviously no surprise that a relational fact cannot be known via a complete description of one of the relata alone (where it is now obvious that by a complete description is meant a complete description of the intrinsic or non-relational properties of the thing). And the reason that this point is irrelevant to the argument against physicalism, forcing at most a minor clarification, is that it seems abundantly clear that having an experience of one phenomenal property rather than another on a given occasion is an intrinsic property of a person, not one that is in any way relational.[\[14\]](#) (To say that such a fact was relational would be to say roughly that it could be altered by altering something about the external relata, while leaving the intrinsic properties of the original thing unchanged.[\[15\]](#))

Second. The other possible rejoinder is an appeal to views about the relations between different "conceptual schemes" or "levels of description" and to related doctrines in the philosophy of science, especially views about reduction. The suggestion, very roughly, is that the Martian scientist might in fact know the very phenomenal facts in question, i.e. that certain propositions within his body of physical and neurophysiological knowledge might describe the *very same facts* that are described by the correct pair of propositions formulated in phenomenal terms, even though the Martian is entirely unable to tell, even in principle, that this is so. And the view which underlies this suggestion is the idea that descriptions of the same fact in different and perhaps incommensurable conceptual schemes need not be logically or analytically or even recognizably equivalent to each other.[\[16\]](#) A full consideration of the complicated issues in the vicinity of this suggestion is obviously impossible within the confines of the present paper, but the following brief remarks may suffice to indicate why I do not find it at all plausible as a response to the present argument.

It is obvious and uncontroversial that particular, concrete entities (objects, states, events) can be picked out or specified in different and not obviously equivalent ways: thus, e.g., Venus as the morning star or as the evening star. It is also obvious that properties can be specified in non-equivalent ways where these specifications are indirect or accidental, i.e. where they pick out the property by invoking a contingent description of it: thus, e.g., one of the phenomenal properties in question might be specified as Joe's favorite color or as the color experienced in connection with a certain standard sort of object. (Indeed, my specifications in this paper were of this sort, which does not of course mean that my own grasp of the properties in question depended on such a specification.) Or, to take a somewhat more interesting case, heat might be specified as the property causally responsible for certain kinds of effects, such as the melting of ice and the cooking of food. But it seems clear that not all property specifications can be thus indirect or accidental, that properties are often specified in a way that reflects or captures their essential or intrinsic character.

And it seems abundantly clear that both the property of having an experience of a certain sensuous color and the various physical and neurophysiological properties, as these are understood by the Martian scientist, are specified in this essential or intrinsic way.[\[17\]](#)

In these terms, the present suggestion is that there could be two different property specifications, each of which captures or represents the essential or intrinsic nature of the very same property rather than picking it out via some sort of indirect or accidental description, but which nonetheless still fail to be logically or even recognizably equivalent to someone who fully understands them both. It seems to me very doubtful that this suggestion is even intelligible, the reason being roughly that properties, unlike most kinds of particulars, simply do not have the right kind of logical depth or complexity to make non-equivalent essential specifications possible. If there are two non-equivalent essential property specifications, I suggest, then there are two properties—however closely related in other ways they may be.

But while I think that the foregoing point is correct as a matter of general metaphysics, I do not want to rely entirely on it here. Thus I propose to grant the physicalist the intelligibility of the present suggestion, at least for the sake of the argument, and see whether it really does him any good. What we are supposing then, applying it to the specific sort of case in question, is that the property specified as being in a certain neurophysiological state and the property specified as having an experience of one of the color properties originally in question are in fact the very same property, even though neither the Martian scientist nor anyone else can tell directly that this is so. But then, as long both specifications are conceded to be intrinsic, it seems to follow that the *single* property in question nonetheless has a kind of internal duality or complexity: it has, we may say, *two* different *aspects* or *dimensions*, one reflected in one specification and one in the other. And now the knowledge that the Martian scientist has no access to will be the knowledge that the latter, experiential aspect or dimension of the property is present on the occasions when the former,

neurophysiological aspect or dimension is. Thus as long as the presence or absence of this experiential aspect or dimension in a particular case is conceded to be a genuine fact, which is something that only the most radical and implausible sort of eliminativism could deny, it will still be the case that the complete physical account leaves out some of the facts and hence, once again, that physicalism is false.[\[18\]](#)

III

My conclusion so far is that the physicalist or materialist view of human mental states is false, on the grounds that certain entirely obvious facts about the qualitative character of phenomenal experience are not captured by any imaginable physical account. I claim no great originality for the argument to this point, for I think that it is very close to what Nagel and especially Jackson had in mind, even though their specific formulations opened the door to irrelevant responses. But unlike Jackson and probably Nagel, I do not think that the force of the argument is restricted to phenomenal experiences, and I will devote the final two sections of the paper to a consideration of how it can be more widely applied, focusing in the present section on states such as propositional attitudes that have intentional content.

Even among those who are doubtful about the case of phenomenal qualia, it has often been supposed that a physicalist account of intentional states like beliefs and desires is on a much sounder footing. This discussion has tended to focus on states of belief, and unfortunately has almost always failed to adequately distinguish the issue of our public belief attribution practices from that of the private or subjective content of the states in question. In the present discussion, I will avoid those complexities by focusing on a simpler sort of state, but one that is still clearly intentional: the state of simply thinking about or envisaging something, of having it in mind.

Suppose then that on a particular occasion I am thinking about a certain species of animal, say dogs—not some specific dog, just dogs in general (but I mean domestic dogs, specifically, not dogs in the generic sense that includes wolves and coyotes). The Martian scientist is present and has his usual complete knowledge of my neurophysiological state. Can he tell on that basis alone what I am thinking about? Can he tell that I am thinking about dogs rather than about cats or radishes or typewriters or free will or nothing at all? It is surely far from obvious how he might do this. My suggestion is that he cannot, that no knowledge of the complexities of my neurophysiological state will enable him to pick out that specific content in the logically tight way required, and hence that physicalism is once again clearly shown to be false.

Before examining this issue, however, it is important to be somewhat clearer than has been necessary so far about the scope of the knowledge that the Martian is allowed to draw on for this purpose. It is natural and, I believe, essentially correct, to regard my having a thought about dogs as a purely internal property of me, one that does not depend in a constitutive way on external objects and situations or on my relations to them (though it may of course be a causal result of such things). This is reflected in the fact that I am able in general to tell “from the inside,” simply by reflection, what I am thinking about, without needing to know anything about these external matters. Thus the Martian should apparently be able to tell on the basis of my internal physiology alone that I am thinking about dogs.

Before arguing specifically that he cannot, I want to consider briefly some possible objections to this construal of the issue, growing out of recent work in the philosophy of language, which challenge the very idea that having a thought with a certain content is an internal property of the person. A full consideration of the various ideas and doctrines involved in these objections is once again obviously impossible within the confines of this

paper. But I believe that it will nonetheless be relatively easy to see that they have no serious effect on the main line of argument being advocated here.

Consider, first, the idea of "the division of linguistic labor." In various papers, Putnam has suggested that I need not have any very clear and determinate conception of, e.g., dogs in order to be thinking about them. It is enough, he seems to suggest, if I merely employ in my thinking the word "dog," with the reference of the word being determined by "the relevant group of experts."[\[19\]](#) Thus, it might be suggested, it would not be at all surprising if the Martian scientist is unable to determine on the basis of my internal neurophysiology alone that I am thinking about dogs, for this fact depends on facts about the experts and not merely on my internal properties.

I think that it is far from obvious that someone who has no conception at all of what sort of thing a dog is, not even that it is an animal as opposed to a vegetable or an inanimate or even an abstract object, is nonetheless thinking about dogs solely by virtue of employing the word. I also doubt that the Martian scientist would find it any easier to determine that I am indeed employing, in the relevant sense, a certain word. But it is enough for present purposes to focus on someone like myself who has a much more detailed conception of a dog. I am not one of the relevant experts (though that would be a possible case too), so it is possible that there are actual non-dogs that I could not distinguish from dogs. And even if I were one of the experts, there would surely be creatures that are at least possible, e.g. perhaps Twin-Earth dogs, that I would be unable to distinguish from real dogs. This is to say that my conception, and probably anyone's conception, of dogs fails to be completely determinate. But this does nothing to solve the main problem, for we can still ask whether the Martian can tell what this somewhat indeterminate thought content is, and the correct answer, I suggest, will still be negative.

Another idea in the same vicinity is the causal theory of reference or perhaps of thought content generally. Again it is suggested that what I am thinking about is not determined by my internal state alone, but depends also on external relations, in this case causal relations, including the causal history of the words I employ. Here too we may concede that there is something right about the point in question. It is at least plausible to think that part of what makes my thoughts pertain to dogs, the earthly species, rather than to Twin-Earth dogs which might be indistinguishable even by the experts at the time in question, is that I am causally related, partly or perhaps even entirely via the causal history of the word, to earthly dogs and not to Twin-Earth dogs. The only thing that we must resist is a *completely* externalist account of content, according to which my internal state possesses by itself no content at all and thus nothing that the Martian could fail to know. And here it is enough, I think, to point out that a completely externalist view of content would be incompatible with the obvious fact pointed out earlier: on a completely externalist view, I would have from the inside no grasp at all of what I was thinking about, since I have in general no access to the relevant causal relations—a result that I take to be obviously and indeed monumentally absurd.[\[20\]](#)

Despite the enormous complexity and subtlety of the recent work in this area, the foregoing is, I think, enough to show that the specific instance of the argument against physicalism that is under discussion in this section cannot be plausibly met by denying that the content of my thoughts is, to a sufficient degree to pose the problem, an internal property of me. Any account of content that makes it accessible enough from the inside to avoid clear absurdity will also make it to that same degree internal, thereby posing a clear challenge to the Martian and hence to physicalism. It may be conceded that there is quite possibly no simple and non-misleading way to specify such purely internal content, thus showing once again the degree to which ordinary language and common sense are insensitive to philosophically significant but practically unimportant (or at least seemingly unimportant)

distinctions. But while this may make the argument somewhat more difficult to formulate, it does nothing at all to affect its basic cogency.[\[21\]](#)

Suppose then, as seems undeniable, that when I am thinking about dogs, my state of mind has a definite internal or intrinsic albeit somewhat indeterminate content, perhaps roughly the idea of a medium-sized hairy animal of a distinctive shape, behaving in characteristic ways. Is there any plausible way in which, contrary to my earlier suggestion, the Martian scientist might come to know this content on the basis of his neurophysiological knowledge of me? As with the earlier instance of the argument, we may set aside issues that are here irrelevant (though they may well have an independent significance of their own) by supposing that the Martian scientist has an independent grasp of a conception of dogs that is essentially the same as mine, so that he is able to formulate to himself, as one possibility among many, that I am thinking about dogs, thus conceived. We may also suppose that he has isolated the particular neurophysiological state that either is or is correlated with my thought about dogs. Is there any way that he can get further than this?

The problem is essentially the same as before. The Martian will know a lot of structural facts about the state in question, together with causal and structural facts about its relations to other such states. But it is clear that the various ingredients of my conception of dogs (such as the ideas of hairiness, of barking, and so on) will not be explicitly present in the neurophysiological account, and extremely implausible to think that they will be definable on the basis of neurophysiological concepts. Thus, it would seem, there is no way that the neurophysiological account can logically compel the conclusion that I am thinking about dogs to the exclusion of other alternatives.

There is, however, one possibility here that is worth brief exploration. A number of philosophers have at least flirted with the idea of what might be called a relational or coherence theory of conceptual content: the idea that concepts are defined entirely by the

formal structure of their inference relations to each other. The further suggestion is then roughly that any system of states that realizes the appropriate formal structure will thereby come to be a genuinely representational system with the concepts in question as the represented content. And if this were so, then the Martian scientist, by knowing the causal structure of my various neurophysiological states, might be able to identify the corresponding contents. (This assumes, obviously and more than a little problematically, that a transition can be made from causal structure to inferential structure, i.e. that causal relations or some appropriately arrived at subset of them can be taken to reflect inference relations.)

There is much that could be said about this sort of picture and a good deal more that would have to be done to make it even minimally plausible. For present purposes, however, two points will suffice. First, even if the coherence theory of concepts is correct, having a structure isomorphic to a given set of concepts will be at most a *necessary*, not a *sufficient* condition for a system of states to actually represent those concepts. There simply is no reason why a system of states could not accidentally happen to have the right structure while in fact representing nothing at all. And thus no structural knowledge on the part of the Martian would show definitively that I was thinking about dogs.

Second, the coherence theory of concepts is in fact very implausible, because it is very implausible that a particular set of concepts can ever be identified on the basis of formal inferential structure alone. On the contrary, there appears to be no reason at all why lots of different sets of concepts could not possess completely parallel and hence indiscernible inferential structures. And this possibility, which is already very serious for concepts considered in the abstract, becomes more serious still when we are dealing with a particular system of concrete states which can plainly never embody all of the possible concepts and inference relations that are abstractly possible, so that two or more systems of concepts that

were abstractly discernible might be equally plausible interpretations of a system of concrete states that did not perfectly embody any of them.[\[22\]](#)

Thus the idea that the Martian scientist would be able to determine the intrinsic or internal contents of my thought on the basis of the structural relations between my neurophysiological states is extremely implausible, and I can think of no other approach to this issue that does any better. The indicated conclusion, once again, is that the physical account leaves out a fundamental aspect of our mental lives, and hence that physicalism is false.

IV

I want to consider one more application of our general line of argument, in some ways the most fundamental of all, but one that is fortunately capable of being dealt with very briefly. It is obvious that on any plausible version of physicalism, only some of our neurophysiological states will be identified with conscious mental states. There is no consciousness associated with those states, for example, that control breathing and heartbeat. But this suggests the issue of whether our Martian scientist, on the basis of his complete physical and neurophysiological knowledge, can tell which neurophysiological states are conscious and which are not. My suggestion, once again, is that there is no way that he can do this in the logically tight way that is required.

We may suppose, reasonably enough, that there is some structural difference between states that are conscious and states that are not, and hence that the Martian can divide our states into two groups, corresponding to this difference. But even if he can get this far, how can he possibly determine, as opposed to merely surmise or conjecture, that the states in one group involve consciousness and that those in the other do not? It is, if anything, even more obvious that consciousness is not explicitly mentioned as such in his complete neurophysiological account, nor definable in terms of things that are mentioned.

And again, as with the case of phenomenal properties, I know of no one who has ever seriously suggested otherwise.

My conclusion, which could, I believe, be extended to many other sorts of mental states as well, is that the Martian scientist, in spite of possessing complete physical and neurophysiological knowledge of me, could not know many important facts about my conscious mental life, nor indeed even that I have a conscious mental life at all. This means that the physical and neurophysiological account is radically incomplete as an account of my complete personal makeup and hence that physicalism or materialism, as an account of human beings, is surely and irredeemably false.

Laurence Bonjour

University of Washington

NOTES

[1] The argument in question may well be a decisive objection to "naturalism" as well, but my understanding of that popular doctrine is too uncertain to warrant very much confidence in such a claim.

[2] Thomas Nagel, "What Is It Like to Be a Bat?" *Philosophical Review*, volume 83 (1974), pp. 435-50; reprinted in David M. Rosenthal (ed.), *The Nature of Mind* (New York: Oxford University Press, 1991), pp. 422-28. References in the text to Nagel are to the pages of this reprint.

[3] Frank Jackson, "Epiphenomenal Qualia," *Philosophical Quarterly*, volume 32 (1982), pp. 127-36. The argument in question is what Jackson calls "the knowledge argument." It receives some useful elaboration in Jackson's note, "What Mary Didn't Know," *Journal of Philosophy*, volume 83 (1986), pp. 291-95; reprinted in Rosenthal (ed.), pp. 292-4. (Subsequent references to this latter article will be to the reprint in Rosenthal.)

[4] See Jackson, "Epiphenomenal Qualia," p. 132, for a bit more discussion of this point. I don't mean to suggest that it is clear that a true physicalist account should not be expected to provide such knowledge, and still less that it is clear why this is supposed to be so. But the issue is difficult at best, so that it is better to find a version of the argument that does not require resolving it.

[5] Jackson, "What Mary Didn't Know," p. 392.

[6] Jackson, "Epiphenomenal Qualia," p. 130.

[7] Paul M. Churchland, "Reduction, Qualia, and the Direct Inspection of Brain States," *Journal of Philosophy*, volume 82 (1985), pp. 8-28. (References in the text to Churchland are to the pages of this paper.)

[8] See Jackson, "What Mary Didn't Know," p. 394.

[9] Jackson, "What Mary Didn't Know," p. 394.

[10] If there is no limit to the levels of detail, if the facts about my states are, as it were, infinitely fine-grained, then the Martian, being finite, does not know absolutely everything about my states. But we may surely stipulate that his knowledge is sufficiently fine-grained to capture everything that is relevant to the issue with which we are concerned here.

[11] One suggestion that has been made in discussion is that it is question-begging against the materialist to assume that it is possible for the Martian to have the same phenomenal experiences even though both his neurophysiological states and their functional roles are presumably different from ours. Actually, it would be quite possible, for all that has been said, to stipulate that the Martian's neurophysiological states are essentially the same as ours, even though hooked up in different ways to sensory mechanisms—and hence different in functional role. Since even many functionalists concede that phenomenal properties are not captured by functional role, this does not seem to beg any serious questions. But the main point is that allowing the Martian to have such phenomenal experiences makes his task easier, not harder, so that it is hard to see on what basis the materialist can object to it. If this is in fact not genuinely possible, then so much the worse for materialism.

[12] Actually it would be somewhat less problematic, and still adequate for my purposes here to suppose merely that the Martian correctly *believes* this to be the case. But it will be simpler and less distracting to speak of knowledge.

[13] Churchland, in his discussion of Mary, suggests as a part of his account of her imaginative extrapolation that color sensations might turn out to be "structured sets of elements" rather than "undifferentiated wholes" [26-7]. I take it that this would mean that color properties were somehow complex, rather than simple, thus at least opening the possibility that they might somehow be definable in terms of neurophysiological primitives. But while it seems clear that something in this direction would be needed to defend the view that the propositions about color experience are indeed logically contained in the neurophysiological account, I can see no real hope that any such view will turn out to be tenable—nor is it clear that even Churchland means to suggest it, given his heavy reliance on the idea that introspective knowledge must first be formulated in neurophysiological terms.

[14] Unless, of course, it involves a relation to a non-physical particular, perhaps a sense-datum or sensum. But it is obvious that this possibility is no help to the physicalist.

[15] David Lewis seems to hold a view according to which the phenomenal character of experience would be in this way relational, by virtue of depending on a choice of an "appropriate" population, in relation to which a person's state is to be classified. See his "Mad Pain and Martian Pain," reprinted in David Rosenthal (ed.), *The Nature of Mind* (Oxford: Oxford University Press, 1991), pp. 229-35. Here I will simply assume that such a view is too implausible to require serious consideration.

[16] See Churchland, *op. cit.*, for one attempt in this direction.

[17] It would in fact be easier to question this claim in the case of the physical and neurophysiological properties, but this would obviously not help the physicalist.

[18] The eliminativist possibility is suggested somewhat obliquely by Churchland in his discussion of Jackson (*ibid.*) and, of course, more explicitly elsewhere. The basic idea that a description in a reducing theory might fail to logically entail a description in a theory being reduced because the theory being reduced is strictly false, albeit close enough to the truth to be regarded as having glimpsed the truth through a glass darkly. Churchland is, of course, quite right that this sort of case is possible in general, as illustrated by various episodes in the history of science. But what is, I believe, too implausible to be taken seriously is the idea that the phenomenal description of my experience is false to the degree that would be required to accommodate in this way the Martian's inability to know which color experience I was having.

[19] See, e.g., Hilary Putnam, "The Meaning of 'Meaning'," in his *Mind, Language and Reality* (Cambridge: Cambridge University Press, 1975).

[20] A philosopher who shall remain nameless once conceded to me in the course of a discussion of this sort of issue that on his view, he could not tell from the inside that we were not discussing quantum mechanics rather than the philosophy of language. It should be easy to see why this made it seem futile to go on with the discussion.

[21] For a somewhat fuller discussion of these recent ideas in the philosophy of language and their bearing on the idea of internally accessible thought content, see my paper "Is Thought a Symbolic Process?" *Synthese*, vol. 89 (1991), pp. 331-52, especially pp. 337-40.

[22] For a somewhat fuller consideration of the coherence theory of content and its implications for thought content in particular, see the paper referred to in the preceding note, pp. 340-45.

