Bealer & Koons (eds), The Waning of Materialism, 2010.

1

Against Materialism

Laurence BonJour

Recent philosophy of mind has been dominated by materialist (or physicalist) views: views that hold that mental states are entirely material or physical in nature, and correlatively that a complete account of the world, one that leaves nothing out, can be given in entirely materialist terms.[1] Though (as the title of this volume suggests) this may be changing to some extent, philosophers of mind who are willing to take seriously the possibility that materialism might be false are still quite rare.

I have always found this situation extremely puzzling. As far as I can see, materialism is a view that has no very compelling argument in its favor and that is confronted with very powerful objections to which nothing even approaching an adequate response has been offered. The central objection, elaborated in various ways below, is that the main materialist view, quite possibly the only serious materialist view, offers no account at all of *consciousness* and seems incapable in principle of doing so. But consciousness, as Nagel pointed out long ago,[2] is the central feature of mental states—or at the very least a feature central enough to make a view that cannot account for it plainly inadequate.

Supposing, as I will try to show below, that this assessment is correct, why have materialist views been so dominant? Part of the answer is that it is far from clear that dualist views, at least those that go much beyond the bare denial of materialism, are in any better shape (see the last section of this chapter for some elaboration of this). But it must be insisted that the inadequacies of dualism do not in themselves constitute a strong case for

[1] Admittedly, the rather stark simplicity of this formulation is not to everyone's taste. There are those who prefer a formulation in terms of the *supervenience* of the mental on the physical, and some (one variant of the recently popular view known as 'non-reductive materialism') who want to interpret such a formulation in a way that gives the mental some sort of ontological independence. (For a recent discussion, see Antony (2007). I have no space here to sort out the twistings and turnings of that discussion and will simply assume that any genuine ontological independence of the mental would amount to a kind of epiphenomenalism (see further below) rather than genuine materialism, and thus that a genuine materialism must be committed to the formulation in the text.

[2] In his classic 1974 paper 'What Is It Like to Be a Bat?' *Philosophical Review* 83:435–50.

materialism: arguments by elimination are always dubious in philosophy, and never more so than here, where the central phenomenon in question (that is, consciousness) is arguably something of which we still have little if any real understanding. Instead, materialism seems to be one of those unfortunate intellectual bandwagons to which philosophy, along with many other disciplines, is so susceptible—on a par with logical behaviorism, phenomenalism, the insistence that all philosophical issues pertain to language, and so many other views that were once widely held and now seem merely foolish. Such a comparison is misleading in one important respect, however: it understates the fervency with which materialist views are often held. In this respect, materialism often more closely resembles a religious conviction—and indeed, as I will suggest further in a couple of places below, defenses of materialism and especially replies to objections often have a distinctively scholastic or theological flavor.

In what follows, I will try to substantiate this indictment of materialism by doing the following things. First, I will look at some of the main considerations that are advanced in favor of materialism in general, as opposed to particular materialist views, attempting to show that these are surprisingly insubstantial and rest mainly on assumptions for which no real defense is offered. Second, I will look at the overwhelmingly dominant materialist view, namely functionalism, arguing that it is deeply inadequate in relation to the problem of consciousness. Third, I will look at what is widely regarded as the most serious specific problem for materialism in general and functionalism in particular, namely the problem of qualitative content or qualia, focusing here on a somewhat modified version of Frank Jackson's well-known "knowledge argument" and trying to show that the objection to materialism that results is still extremely compelling. Fourth, I will look at a problem that functionalism is claimed to handle more successfully, the problem of intentional states, arguing that there are clear cases of *conscious* intentional states which materialism in general and functionalism in particular can handle no better than qualia—and for essentially the same reasons. Fifth, and last, I will ask what lessons, if any, for a more adequate account of conscious mental states can be derived from all of this.

1. THE CASE FOR MATERIALISM

One of the oddest things about discussions of materialism is the way in which the conviction that *some* materialist view must be correct seems to float free of the defense of any particular materialist view. It is very easy to find people who seem to be saying that while there are admittedly serious problems with all of the specific materialist views, it is still reasonable to presume that *some* materialist view must be correct, even if we don't know yet

which one, or that the seeming force of the objections to particular materialist views must be balanced against the strength of the underlying case for materialism. But why is this supposed to be a reasonable stance to take? What arguments or reasons or otherwise compelling intellectual considerations are there that could yield a strong background presumption of this sort in favor of materialism (or create a substantial burden of proof for opponents of materialism)?

There are, of course, arguments *against* particular versions of dualism, mainly against the interactionist version of Cartesian substance dualism. For reasons already mentioned, I will set these aside as not constituting in themselves an argument *for* materialism. There is also the inductive generalization from the conspicuous success of materialist science in a wide variety of other areas. This undeniably has some modest weight, but seems obviously very far from being enough to justify the strong presumption in question. Inductions are always questionable when the conclusion extends to cases that are significantly different from the ones to which the evidence pertains, and even most materialists will concede that conscious phenomena are among the most difficult—indeed, seemingly the most difficult of all—for materialist views to handle. Thus the fact that materialism has been successful in many other areas does not yield a very strong case that it will succeed in the specific area that we are concerned with.

Beyond this, there seem to be only two related sorts of grounds that are offered for a strongly pro-materialist presumption, both of which are quite flimsy, when subjected to any real scrutiny.

1.1. The 'Principle' of Causal Closure

The first and clearer of these two grounds appeals to the thesis that the material universe is *causally closed*: that material things are never causally affected by anything non-material (so that, as it is often put, physical science can in principle give a completely adequate explanation of any physical occurrence, without needing to mention anything non-physical). This thesis is commonly referred to as a "principle," a characterization that leaves its status rather obscure. (Philosophers often seem to describe something as a "principle" when they are inviting their readers to accept it as a basis for further argument, even though no clear defense of it has been offered.)

The closure principle does not by itself entail that materialism is true. It leaves open both the possibility of non-material realms that are causally isolated from the material world and also the possibility that epiphenomenalism is true: that conscious phenomena are side-effects of material processes that are incapable of having any reciprocal influence on the material world. But, assuming that the non-material realm in the first possibility is supposed to be the locus of conscious phenomena, both of these possibilities are extremely unpalatable, even paradoxical, in essentially the same way. The main problem is not, as

is often suggested, that they are incompatible with the general common sense intuition that conscious states causally affect bodily behavior. A more specific and serious problem is that if either of these possibilities holds, then it becomes difficult or seemingly impossible to see how verbal discussions of conscious phenomena—such as this chapter and many others—can be genuinely about them in the way that they seem obviously to be. How can people be talking about conscious states or saying anything significant about them if completely adequate causal explanations of their verbal behavior can be given that make no reference to such states? Even without invoking any specific version of the causal theory of reference, it is hard to see why verbal discussions that are entirely unaffected by what they purport to be about should be taken seriously. Thus while a number of philosophers have in recent times been seemingly tempted by epiphenomenalism, it appears that they can have been genuinely advocating such a view about conscious states only if the view itself is false.[3]

For these reasons, the argument from the principle of causal closure to the truth of materialism is quite strong, even if not fully conclusive. But why is the principle of causal closure itself supposed to be so obviously correct? Clearly this 'principle' is not and could not be an empirical result: no empirical investigation that is at all feasible (practically or morally) could ever establish that human bodies, the most likely locus of such external influence, are in fact never affected, even in small and subtle ways, by non-material causes. We are told that scientists accept this principle, and often that most philosophers accept it as well. But do they have any compelling reasons for such acceptance? Or is this vaunted principle nothing more than an unargued and undefended assumption—a kind of intellectual prejudice, in the literal meaning of the word?

Taken in the abstract, apart from any appeal to a specific account of conscious mental phenomena, I have no idea whether the principle of causal closure is true or not. More importantly, I cannot imagine how to rationally decide whether it is true without *first* arriving at a defensible account of conscious mental states. It seems utterly obvious that mental states do causally affect the material realm: probably by causally affecting the actions of human bodies in general, but (as just argued) at least more narrowly by causally affecting verbal discussions of these matters. *If* a materialist account of conscious states is correct, then the principle of causal closure seems likely to be true. But if no such account is correct, then the principle is almost certainly false. Thus to argue for the truth of materialism or for a strong presumption in favor of materialism by appeal to the principle of causal closure is putting the cart in quite a flagrant way before the horse.

[3] This problem seems to be the main reason for Jackson's abandonment of his previous anti-materialist stance. (Jackson never took seriously the possibility that the non-material qualia for which he was arguing might causally affect the material world.)

1.2. The Appeal to 'Naturalism'

A second sort of defense of a general presumption in favor of materialism appeals to the general idea of *naturalism*. Here again we have a view, like materialism itself, to which many, many philosophers pay allegiance while offering little by way of clear argument or defense, but here the view itself is much harder to pin down in a precise way. Indeed, even more striking than the absence of any very clear arguments is the fact that many recent philosophers seem so eager to commit themselves to naturalism—to fly the naturalist flag, as it were—while showing little agreement as to what exactly such a commitment involves. Thus naturalism seems to be even more obviously an intellectual bandwagon than materialism. (In addition, naturalism, for some of those who use the term, seems to just amount to materialism, which would make an argument from naturalism to materialism entirely question-begging.)

Is there any genuine support for a materialist presumption to be found in the vicinity of naturalism? One version of naturalism is the idea that metaphysical issues—or philosophical issues generally—should be dealt with through the use of the methods of natural science. If this is accepted, and if it is true that following the methods of natural science leads plausibly to an endorsement of materialism, then at least some presumption in favor of materialism might follow. But both of the needed suppositions are in fact extremely dubious, to say the least. There is simply no good reason to think that the methods of natural science exhaust the methods of reasonable inquiry—indeed, as has often been pointed out, there is no plausible way in which that claim itself can be arrived at using those methods. Nor is there any very clear reason to think that applying the methods of natural science to the question of whether materialism is true, assuming that one could figure out some reasonably clear way to do that, would lead to the conclusion that materialism is correct. Such a conclusion is obviously not within the purview of physics, but it is also not within the purview of psychology, especially as currently practiced. As was true with closure, there is no doubt that many (but not all) natural scientists *assume* the truth of materialism, but the question is whether they have any good reason for such an assumption—a reason that would itself have to transcend their strictly scientific claims and competence.[4]

Thus, while the murkiness of the discussions of naturalism makes it harder to be sure, naturalism, like closure, does not seem to yield an independently defensible presumption in favor of the truth of materialism. If there is any better

[4] Lurking here is the difficult issue of what sorts of entities or properties count as material or physical. Is there any good way to delimit the realm of the material that does not preclude further discoveries in physics, but also does not trivialize the category by allowing it to include anything that people in departments labeled "Physics" might eventually come to study? This is anything but a trivial problem, but I have no space here to pursue it further.

reason or basis for such a presumption that is prior to and independent of the defense of some particular materialist view, I have no idea what it might be.

2. FUNCTIONALISM AND CONSCIOUSNESS

The upshot of the previous section is that the case for materialism must rest almost entirely on the defense of particular materialist views and not to any substantial extent on any background presumption. So what materialist views are there? The answer, I think, is that once both logical behaviorism and various versions of eliminativism are set aside as too implausible to be taken seriously—something that I will assume here without any further discussion—there is only *one* main materialist view, namely functionalism, with no very serious prospect that any others will emerge. And the fundamental problem for materialism, I will suggest, is that functionalism offers no account at all of consciousness and seems in principle unable to do so.

What gives rise to the mind–body problem in the first place and poses the essential problem that any adequate version of materialism must solve is the fact that conscious mental states, as we are aware of them, do not present a material appearance—do not seem as we experience them to be material in their makeup in any apparent way. Thus a view which holds that everything that exists is material must either (a) deny the very existence of such states, as eliminativism does, or else (b) explain how states and correlative properties that do not initially seem to be material in nature can nevertheless turn out to be so. A view that takes the latter alternative must give an account of the nature of such states and properties that both accurately reflects their character as experienced and explains how they can nonetheless be entirely material in their makeup. And this, I suggest, is something that has never been successfully done.

The starting point for modern versions of materialism was the central-state identity theory, particularly the version advocated in a famous paper by J. J. C. Smart (1959). Smart recognized that the truth of materialism can only be an empirical discovery, not something knowable a priori. For this to be so, he argued, the various mental states in question must be conceived in a *topic-neutral* way: a way that makes it *possible* for them to be merely material in character, without implausibly *requiring* that this be so. Only in relation to such a conception would it be possible to discover empirically that such a state is in fact a neurophysiological state of some kind.

But for this to work, it is crucially important that the topic-neutral conception in question be adequate to capture the essential features of mental states—something about which Smart was less clear than he might have been. For only if this is so will it be the case that showing that the conception offered can be realized by a material state can establish that mental states might in fact be merely material states, thereby allowing the rest of the argument to proceed on grounds of simplicity, as Smart suggests. Whereas if the proposed topic-neutral conception leaves out essential features of mental states—such as consciousness—then the fact that material states can satisfy that conception will be insufficient to explain how mental states might just be material states. (Smart's own attempt at a topic-neutral characterization fails to distinguish conscious mental states from whatever else might be "going on" in the person under a particular set of circumstances.)

As in Smart's view, functionalism in effect attempts to offer a topic-neutral characterization of mental states, one which will allow but not require that they be essentially material in character.[5] The more general functionalist characterization is in terms of *causal role*: a mental state is characterized by its causal relations to sensory inputs, behavioral outputs, and other mental states of the same sort. The functionalist then proceeds to argue that the states thus characterized could perfectly well be material states, even though the functional characterization does not require this. A further, widely discussed, aspect of the view is that different material states could satisfy the functionalist characterization of a particular mental state in different sorts of creatures or even in the same creature at different times, so that (on the most standard version) a material state *realizes* a functionally characterized state but is not strictly *identical* with it.

But the deepest problem for the functionalist is that the characterization of mental states in terms of causal role says *nothing at all* about consciousness or conscious character. There is no apparent reason why a state that realizes a particular causal role would thereby need to have any specific sort of conscious character (the point made by the familiar reversed spectrum cases)—or indeed any conscious character at all. Thus to point out that a physical state could realize such a causal role really does *nothing at all* to explain how a conscious state could be (or be realized by) a merely physical state. In this way, functionalism fails utterly to offer any explanation or account of the most important and conspicuous feature of mental states—or, at the very least, of a very important and conspicuous feature.

It may seem hard to believe that a view that has been held by so many people for so long can be so easily shown to be inadequate in a fundamental way, but I think that this is nonetheless so. The only solution would be

[5] This way of looking at functionalism is explicit in David Lewis's discussion of one of the earliest versions of the view in his 1966 paper 'An Argument for the Identity Theory,' *Journal of Philosophy* 63: 17–25, see p. 20.

to offer some supplementary account of what material features give rise to conscious experience. But I know of no such account, at least none with any real plausibility.[6]

This difficulty with materialism in general and functionalism in particular has of course occasionally been recognized.[7] But it still seems to have had remarkably little impact on the widespread acceptance of materialist views. I have no very good explanation to offer of this, though part of the reason is perhaps the prevailing tendency to approach the philosophy of mind from a third-person, neo-behaviorist perspective, in which consciousness is largely or entirely ignored. (But on this issue, it is hard to distinguish the chickens from the eggs.)

My basic case against materialism is complete at this point: there is no good reason for any strong presumption in favor of materialism; and the main materialist view fails to offer any real explanation of a central aspect of mental states, namely their conscious character, meaning that there is no good reason to think that it is correct as an account of such states. But though this very simple argument seems to me entirely compelling, I will elaborate it further in the next two sections by focusing on the two main specific kinds of mental states. The version of the argument that applies to states with qualitative content is very familiar, even though I think that its full force has still not been generally appreciated. In contrast, the application of essentially the same basic argument to conscious states with intentional content has received far less attention.

3. THE PROBLEM OF QUALITATIVE CONSCIOUSNESS: MARY REDUX

Though functionalism fails to adequately account for consciousness of any sort, perhaps the most conspicuous aspect of this failure pertains to qualitative content: the sort of content involved in experiences of color and sound, and of things like pains and itches. This point has been made in many ways, but the most straightforward and compelling in my view is still the so-called "knowledge argument," initially suggested by Thomas Nagel in relation to the experiences of bats and later developed by Frank Jackson using his famous example of black-and-white Mary, on which I will mainly focus here. (As most will know,

[6] One possibility is the so-called higher order thought theory, which holds that consciousness arises when one mental state is the object of a second, higher order mental state. I have no room here to consider this view in detail. But the basic—and obvious—problem with such a view is that there is no reason why there could not be a hierarchy of sort, even one with many more levels, in which there was no consciousness involved at all. (For some elaboration, see my contribution to BonJour and Sosa (2003), pp. 65–8.)

[7] See, for example, Colin McGinn (1989); David Chalmers (1995); and David Chalmers (1996).

Jackson has changed his mind about this argument and now rejects it, though his reasons seem to me unpersuasive.)

I will assume here that Jackson's original version of the saga of Mary is familiar enough to require only a brief summation. Mary is a brilliant neurophysiologist, 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 visual perception, by knowing their causal relations to sensory inputs, behavioral outputs, and other such states. But despite all of this knowledge, Mary apparently 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, and analogous things about other qualitative experiences. '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.'[8]

Despite the initial force of this rather simple argument, materialists have not been persuaded, and the literature comprising materialist responses to the Mary example is very large.[9] One thing to say about these responses is that few if any of them are even claimed to have any substantial independent plausibility; instead they are put forward in a way that takes for granted the sort of general presumption in favor of materialism and correlative burden of proof for anti-materialist views that I have argued does not genuinely exist. A full discussion of these responses is impossible here, but there are some main themes that can be usefully dealt with in a general way.[10] One of these is the suggestion that although Mary undeniably acquires something new when she leaves the black-and-white room, what she acquires is not a knowledge of a new *fact* (or facts), but rather something else. A second is the suggestion that what she does acquire is instead something like a new *ability*, perhaps more specifically a new conceptual or representational ability. And if these two themes are combined, it is claimed, the

[8] Frank Jackson (1982), p. 130; see also Frank Jackson (1986).

[9] Many of these discussions are collected in Ludlow, Nagasawa, and Stoljar (2004).

[10] For a useful taxonomy of the various possible materialist responses, see Robert Van Gulick, 'So Many Ways of Saying No to Mary,' in Ludlow, Nagasawa, and Stoljar (2004), pp. 365–405. (This is of the places where the materialist discussion bears a striking similarity to scholastic theology: one can easily imagine a complacent theist writing an article entitled 'So Many Ways to Answer the Problem of Evil.')

result is that there is nothing about the Mary example that is incompatible with the truth of materialism.

Particularly in light of the general materialist failure to provide an account of conscious experience, I doubt very much whether any response of this sort would seem even mildly convincing to anyone who was not already determined to adhere to materialism come what may. But the first of these two themes does at least point to a kind of lacuna in Jackson's original account of the case: if Mary learns new facts, what exactly are they? Indeed, in his response to an early version of this suggestion, Jackson is reduced to invoking the problem of other minds as a (very!) indirect basis for thinking that factual knowledge of some sort is involved.[11]

It is, however, surprisingly easy to modify the original case in a way that makes it utterly clear that there are facts that Mary does not know while she is in the black-and-white room and will learn when she emerges. Suppose that while she is still in the otherwise black-and-white environment, two color samples are brought in: one a sample of a fairly bright green, approximately the color of newly mown grass, and the other a sample of a fairly bright red, approximately the color of a fire engine. Mary is allowed to view these samples and even to know that they are two of the 'colors' that she has learned about in her black-and-white education. She is not, however, told the standard names of these colors, nor is she allowed to monitor her own neurophysiology as she views them.

We now remind Mary of two specific cases that she has studied thoroughly and about which she knows all the physical/neurophysiological/functional facts. One of these is a case where a person was looking at newly mown grass, and the second is a case where a person was looking at a newly painted fire engine. We tell Mary that one of these people had an experience predominantly involving one of the colors with which she is now familiar and that the other person had an experience predominantly involving the other color, but of course not which was which. If we call the colors presented by the samples *color A* and *color B*, Mary now knows that one of the two following pairs of claims is true:

(1) The experience of freshly mown grass predominantly involves color A, and the experience of a newly painted fire engine predominantly involves color B.

(2) The experience of freshly mown grass predominantly involves color B, and the experience of a newly painted fire engine predominantly involves color A.

But can she tell, on the basis of her black-and-white knowledge, together with her new familiarity with the two colors, whether it is (1) or (2) which is true? (Notice carefully that there is no apparent problem with her ***understanding*** of these claims.)

[11] See Jackson (1986: 294).

Though we have made things vastly easier for Mary by focusing on two cases involving colors with which she is now familiar, rather than asking her to figure out on the basis of her overall physical/neurophysiological/functional knowledge what color experiences in general are like, it still seems quite clear, for essentially the same reasons that were operative in the original case, that she will have no more success with her much more limited task. Just as there was nothing in the physical account that could tell her what an experience of red was like, so there is still nothing in the physical account of the fire engine case that could definitively pick out the color of one of the samples as opposed to the other.[12] And yet whichever of (1) and (2) is true states a *fact* (or facts) in as robust a sense as one could want—a fact that Mary will learn when she emerges from the black-and-white room and is allowed to view ordinary objects of various kinds.

Moreover, if there are *abilities* that result from experiencing the two colors in question, Mary presumably can acquire them on the basis of the samples. Consider, for example, Harman's suggestion[13] that what Mary acquires in the original case, when she leaves the black-and-white room and sees red for the first time, is a *perceptual concept* of red, one that essentially involves being disposed to form perceptual representations involving it in the presence of causal stimulation of the right sort—so that she cannot acquire *that* concept in the original version of the black-and-white room. There is much that is questionable about the idea of such a concept, but if there is indeed such a thing, then Mary in the new version of the case presumably can acquire it by viewing the red sample. (Perhaps more than one sample is for some reason required, but it would be easy enough to modify the new version of the case to allow for that.) So, we may suppose, Mary has the perceptual concept of red and the perceptual concept of green, but she still cannot figure out from her physical knowledge which of these concepts is being employed by the people in the cases she has studied. Yet this too is a fact, and if materialism is true, an entirely physical fact. So why can't she know it?

Here, as far as I can see, there are only two possible moves for the materialist which are even marginally worth considering. One is the suggestion that Mary *already* knows the facts in question, as a part of her overall physical knowledge, but that she knows them under a different 'guise' or 'mode of presentation' than that under which she will come to know them when she leaves the black-and-white room. This idea can be developed in different ways and with enormous technical ingenuity. But does it really have any serious plausibility? Imagine that

[12] As Jackson emphasizes in Jackson (1986: 295), it is not enough for Mary to be able to conjecture or guess at the answer to this question. For physicalism to be true, the fact in question must actually be *contained* in her physical knowledge.

[13] Gilbert Harman (1990), 'The Intrinsic Quality of Experience,' *Philosophical Perspectives* 4: 31–52, at pp. 44–5. Harman does not actually mention the Mary case as such, focusing instead on a person who is blind from birth but still learns 'all the physical and functional facts of color perception.' But he does cite Jackson (along with Nagel) as the source of the objection he is discussing.

Mary, in our modified version of the case, having finally experienced real colors, is eager to find out more about these intriguing features of the world about which she has been kept in ignorance. She wants very much, for example, to know whether it is (1) or (2) that captures the relevant facts about cases of that sort—and is seriously frustrated about being kept in ignorance any longer. Suppose that we respond to her frustration by informing her that she already knows the very facts that she is so eager to learn. Surely she would not be satisfied. How might she respond?

I think we can imagine Mary saying something like this:

> You philosophers are really amazing! The idea that I *already* know the facts I am interested in—indeed all facts of that general kind—is simply preposterous. I know all of the physical details, but none of them tells me which of the properties I have just experienced, on the basis of the samples, is realized in each of the two cases. If you suggest to me that there aren't really novel properties, but rather novel concepts or ways of representing or whatever, then (while finding that suggestion itself pretty hard to swallow) I would still insist that *which* concept or way of representing is involved in each case is still something that my physical knowledge doesn't give me any clue about. Perhaps, as you say, there is some clever or complicated way in which the things I want to know are related to the physical things I do know—maybe there is even some metaphysically necessary connection between them (assuming that it is kosher for materialists to believe in such things!). Anything like that, however, just *adds* to the list of facts that my physical knowledge doesn't reveal to me. I am a scientist and not a philosopher, so I'm not really sure which conception of a fact is the right one. (All of the ones you suggest seem pretty weird.) But there is undeniably something that I want to know—something that is true about the world—that can't be learned on the basis of all my physical knowledge. And that means that the physical story isn't in fact the whole story!

Not surprisingly, I think that the response I have imagined for Mary is exactly right—that any way of understanding or individuating facts according to which some piece of Mary's physical knowledge and either (1) or (2) above turn out to be formulations of the same fact is a conception of fact that is simply too intuitively implausible to be taken seriously.

The other possible materialist response is to grant that Mary will learn new facts when she emerges from the black-and-white room in the modified version of the case, but to insist that these are nonetheless still *physical* facts. On this view, what the case shows is that it is impossible for Mary to acquire complete physical knowledge in the black-and-white room. One way to put it is to say that while she can learn all the *objective* physical facts, there are still certain *subjective* physical facts[14] that she can't learn. One can learn what it feels like subjectively to be an organism of such-and-such a general physical description in such-and-such a specific physical state only by actually realizing that condition. But that it feels

[14] See Van Gulick (2004) for one version of this suggestion.

a certain way or involves a certain sort of conscious experience is still, on this view, an entirely physical fact.

I have to admit that I find it nearly impossible to take this response seriously. The only argument for it seems to be an appeal to a background presumption in favor of materialism that is so strong as to make it allegedly reasonable to claim that any fact there is *must* be a material fact, even if we can't see in any clear way how it could be a material fact. Materialism, as we have already seen, offers no real account or explanation of consciousness and so also no reason for thinking that there is any subjective experience at all involved in being in a certain material state. Thus to advance a view of this sort is in effect just to insist that no fact of any sort can be allowed to refute materialism and thus that any possibility of this sort must simply be absorbed into the materialist view, however inexplicable in materialist terms it may be. (It is not much of a stretch to imagine the materialist saying that we must first believe in order than we may understand.)

Thus the modified version of the Mary case seems to present an objection to materialism in general (and functionalism in particular) that is about as conclusive as philosophical arguments ever get. However exactly they should be characterized, there are facts that Mary cannot know on the basis of her complete physical/neurophysiological/functional knowledge, even when she is given the sort of limited experience needed to understand the claims in question and to acquire any abilities that might be relevant. These facts do not seem to be material facts, and there is no basis that is not utterly arbitrary and question-begging for supposing that they are. Thus we have the strongest of reasons for holding that the materialist account of reality is incomplete and hence that materialism is false.

4. THE PROBLEM OF CONSCIOUS INTENTIONAL CONTENT

Qualia of the sort involved in the Mary case are widely recognized to pose a serious problem for materialist views, and it is not too hard to discern occasional misgivings in this respect under the façade of materialist confidence. But, as already mentioned above, it is widely assumed that materialism is in much better shape with regard to intentional mental states: propositional attitudes and other states that involve "aboutness." I believe, however, that this is almost entirely an illusion—that the problems for materialism are just as serious in this area, with consciousness being once again the central focus.

Materialist accounts of intentional states tend to focus mainly on dispositional states, such as beliefs and desires. Given the central role of such states in explanations of behavior, this is in some ways reasonable enough. But such a focus tends to neglect or even ignore the existence of *conscious* intentional states—even though having conscious thoughts that P is surely one of the

central things that having a dispositional belief that *P* disposes one to do. A focus on belief in particular also has the unfortunate effect of making externalist accounts of intentional content seem more plausible than they possibly could if the emphasis were on conscious intentional states.

For these reasons, I will focus here explicitly on conscious thoughts. As I sit writing this chapter, a variety of conscious thoughts pass through my mind. Many of these involve the assertion or endorsement of various propositions: that materialism cannot account for consciousness, that the trees outside my window are very bare, that the weather looks cold and dank, that the situation in the Middle East looks grim, and so on. Other thoughts are also propositional, but in a way that does not involve assertion: my conscious desire to get several pages written before lunch, the hope that the stock market will continue to rise, and so on. It is doubtful that conscious thought must always be propositional in character, but it will in any case simplify the issues to be discussed if we largely ignore the propositional aspect of these various thoughts and focus simply on their being conscious thoughts *of* or *about* various things or kinds of things: materialism, the trees, the Middle East, the stock market, and so on.

One crucial feature of such conscious thoughts is that when I have them, I am in general consciously *aware of* or consciously *understand* or *grasp* what it is that I am thinking about (and also what I am thinking about it). When I think that the trees outside my window are bare, I consciously understand that it is certain *trees* that I am thinking about (and along with this, what sort of thing a *tree* is, and *which* trees I have in mind). What exactly this conscious grasp of the object of thought involves varies from case to case and is sometimes not easy to precisely specify. Moreover, as will emerge, it is something of which I think we presently have no real explanatory account of any substance. But its existence is, I submit, completely undeniable. Indeed, being able in this way to consciously think about things, to have them in mind, is in many ways the most central and obvious feature of our mental lives.

It is obvious that a person's conscious grasp of the object of their thought, of what they are thinking about, can vary on a number of dimensions: it may be more or less precise, more or less detailed, more or less clear, more or less complete. But contrary to what is sometimes suggested, it is rarely if ever merely *disquotational* in character. *Perhaps* (though I doubt it) there are cases where a scientifically untutored person is thinking about, e.g., electrons, and where their sole grasp of what they are thinking is that it is what is referred to in their society or community by the word 'electron'—so that what they are thinking about is in effect: "electrons" (whatever *they* are). But this is surely not the ordinary situation when we think about various things.[15]

[15] Notice that even a thought about what the relevant *societal experts* mean by "electrons" would have to involve a non-disquotational element in the reference to those experts and also in the reference to the word: to think about 'whatever it is that the societal experts mean by "electrons"' is not the same thing as thinking about 'whatever is meant by "electrons" by whatever is meant by "the societal experts."'

Moreover, the existence of conscious intentional content is perfectly compatible with the existence of an externalist dimension of thought content—though not with the view that *all* content is external. If, as in Putnam's famous example, a person is thinking about earthly water at a time prior to the discovery of its chemical composition, there is no reason to deny that they are, in a sense, thinking about H_2O. But in such a situation, the aspect of being about H_2O will obviously not be part of their conscious, internal grasp of what they are thinking about in the way that the more superficial aspects of water will.[16] And in a somewhat parallel way, the person in Burge's famous example who thinks he has arthritis in his thigh and to whom our standard belief-ascription practices ascribe a belief about *arthritis* (where this is, among other things, a disease that only occurs in joints) obviously does not consciously grasp the disease that he is thinking about in a way that involves this specific feature of it.[17] But it is nonetheless impossible to describe either example in a convincing way without presupposing that the people in question do have *something* consciously in mind: a substance having the superficial properties of water in Putnam's example; and a disease having certain fairly specific features in Burge's. Thus while it is possible to dispute the relative importance of conscious, internal thought content and external thought content of which the subject is not conscious, examples of this sort provide no basis at all for denying that conscious internal content exists.

The issue I want to raise here is whether a materialist view can account for the sort of conscious intentional content just characterized. Can it account for conscious thoughts being about various things in a way that can be grasped or understood by the person in question? In a way, the answer has already been given. Since materialist views really take no account at all of consciousness, they obviously offer no account of this particular aspect of it. But investigating this narrower aspect of the issue can still help to deepen the basic objection to materialism.

Here it will be useful to bring the brilliant neurophysiologist Mary briefly back onto the scene, even though the black-and-white aspect of her situation is no longer relevant. Suppose that Mary studies me as a subject and comes to have a complete knowledge of my physical and neurophysiological makeup as I am thinking these various thoughts. Can she determine on that basis what I am consciously thinking about at a particular moment?

One thing that seems utterly clear is that she could not do this merely on the basis of knowing my *internal* physical characteristics—as it is sometimes put, knowing everything physical that happens inside my skin. There is no reason

[16] See Hilary Putnam (1975a). (Cambridge: Cambridge University Press), pp. 215–71.

[17] See Tyler Burge (1979).

at all to think that the internal structure of my physical and neurophysiological states could somehow by itself determine that I am thinking about weather rather than about the Middle East or the stock market.

A functionalist would no doubt say that it is no surprise that Mary could not do this. In order to know the complete causal or functional role of my internal states, Mary also needs to know about their external causal relations to various things. And, it might be suggested, if Mary knows all of the external causal relations in which my various states stand, she will in fact be able to figure out what I am consciously thinking about at any particular time. No doubt the details that pick out any particular object of thought will be very complicated, but there is, it might be claimed, no reason to doubt that in principle she could do this.

Here we have a piece of materialist doctrine that again has a status very similar to that of a claim of theology. It is obvious that no one has even the beginnings of an idea of how to actually carry out an investigation that would yield a result of this kind—that the *only* reason for thinking that this could be done is the overriding assumption, for which we have found no cogent basis, that materialism *must* be true. Among a multitude of other difficulties, Mary would have to be able to figure out the content of thoughts that are confused or inaccurate, or thoughts about imaginary or fictional entities or supernatural entities. It is, to say the least, *very* hard to see how she could do this on the basis of a knowledge of causal relations to more ordinary sorts of things.

But the problem for materialism is in fact even worse than that. For, as already emphasized, it is an undeniable fact about conscious intentional content that I am able for the most part to consciously understand or be aware of what I am thinking about 'from the inside.' Clearly *I* do not in general do this on the basis of external causal knowledge: I do not have such knowledge and would not know what to do about it if I did. All that I normally have any sort of direct access to, if materialism is true, is my own internal physical and physiological states, and thus my conscious understanding of what I am thinking about at a particular moment must be somehow a feature or result of those internal states alone. Causal relations to external things may help to *produce* the relevant features of the internal states in question, but there is no apparent way in which such external relations can somehow be partly *constitutive* of the fact that my conscious thoughts are about various things in a way of which I can be immediately aware. But if these internal states are sufficient to fix the object of my thought in a way that is accessible to my understanding or awareness, then knowing about those internal states should be sufficient for Mary as well, without any knowledge of the external causal relations. And yet, as we have already seen, it seems obvious that this is not the case.[18]

[18] It is worth noting that the same thing is really true in the case of qualia as well. A person's awareness of one color rather than another when he or she looks at newly mown grass obviously does not depend in a constitutive way on external causal relations, even though it may be causally produced by them. Thus in that case too, a knowledge of the person's internal physical and neurophysiological states alone should enable Mary to pick out one color rather than the other as the right one. But it is even more obvious than in the original case that this is not so.

Thus we have the basis for an argument that is parallel to Jackson's original argument about qualia: Mary knows all the relevant physical facts; she is not able on the basis of this knowledge to know what I am consciously thinking about at a particular moment; but what I am thinking about at that moment is as surely a fact about the world as anything else; therefore, complete physical knowledge is not complete knowledge, and so materialism is false.

One way to further elaborate this point is to consider how it applies to what is perhaps the most widely held materialist view of intentional content: the view, popularized by Jerry Fodor and many others, that intentional mental states employ an internal *language*, a "language of thought."[19] Fodor calls this view 'the representational theory of the mind,' though it might better be called 'the symbolic theory of the mind.' For the crucial feature of the view is that the language of thought, like any language, is composed of *symbols*: items that do not stand for anything by virtue of their intrinsic properties, but whose representative character depends instead on the *relations* in which they stand to other things—for Fodor, the sorts of causal relations that are captured in the idea of a causal or functional role.[20] Just as the word "dog" could in principle have stood for anything (or nothing at all) and in fact stands for a kind of animal rather than something else only because of causal relations that arise from the way it is used, so also the symbols in the language of thought stand for whatever they stand for only by virtue of analogous sorts of relations and not in virtue of their intrinsic physical and neurophysiological properties. Their intentional character is thus *extrinsic*, not *intrinsic*.

Proponents of the language of thought rarely have much to say about conscious thoughts of the sort that we are focusing on here. But it is clear that on their view, what happens when I am consciously thinking about, say, the Middle East is that in some appropriate location in my overall cognitive operations there occurs a symbol (or set of symbols) that refers to the Middle East. This symbol, like the surrounding context in which it occurs, is some neurophysiological state or some constellation of such states. No one, of course, has at present any real knowledge of the concrete nature of such symbols or their larger contexts, but it will do no harm to follow Fodor in thinking of a mental "blackboard" on which mental symbols are inscribed in appropriate ways. Thus for me to be consciously thinking about the Middle East is for me to have the mental symbol that refers to

[19] See, e.g., Jerry Fodor, 'Propositional Attitudes' and 'Methodological Solipsism Considered as a Research Strategy in Cognitive Science,' both reprinted in his 1981 book *Representations* (Cambridge, MA: MIT Press).

[20] See, for example, Jerry Fodor (1987), chapter 4. Fodor has subsequently refined this view in various ways, but none that affect the issues being raised here.

the Middle East inscribed in the right way on this "blackboard." But the symbol's reference to the Middle East, to repeat, depends not on its intrinsic physical or neurophysiological character alone, but also on the relations in which it stands to other such symbols and, directly or indirectly, to the external world.

Suppose now that Mary is studying my cognitive operations. Suppose that she has somehow isolated what amounts to my mental "blackboard" and the various symbols "written" on it. Obviously this will not in itself tell her what I am thinking about. Even if she could somehow focus on the specific symbol that refers to the Middle East and tell that it is functioning in a way that determines the object of my conscious thought (even though there is no reason to think that she could in fact do these things), she will not on this basis alone be able to tell what it is that this symbol in fact refers to. Nor is there any plausibility to the idea that Mary could figure out the reference or meaning of the various mental symbols simply by examining their internal relations to each other.[21] Thus she will need once again to appeal to external causal relations of various sorts.

But how then am I able to be aware of or understand "from the inside" what I am thinking about? Once again I have no knowledge of those external relations (and would be very unlikely to be able to figure anything out from them even if I did). All that I plausibly have access to is the mental symbol or symbols and the surrounding system of states, and this is apparently not enough to determine the object of my thought.

The only very obvious recourse here for the proponent of a language of thought is to construe my understanding or awareness of what I am thinking about disquotationally in relation to the language of thought. When thinking about the Middle East, I do so by using some mental symbol. And when I understand or am aware of what I am thinking about, it might be suggested, I in effect use that very same symbol: what I am aware of is that I am thinking about 'the Middle East' (whatever that is—that is whatever that symbol in fact refers to). If the symbol in question did succeed in referring to the Middle East, then this specification of what I am thinking about will refer to the Middle East as well and so will be correct. But it is intuitively as obvious as anything could be that my awareness of what I am thinking about normally involves more than this: involves actually understanding (at some level of precision, detail, etc.) what the Middle East is in a way that goes beyond merely repeating the same symbol. Assuming for the moment that there really is a language of thought, I *understand* my language of thought in a way parallel to the way in which I understand my own public language—and not in the merely disquotational way that could just as well be applied to a language of which I have no understanding at all.

Here a proponent of the language of thought may want to reply that the difference in the public language case is merely that one language is a language

[21] Such an idea has sometimes at least apparently been suggested. For more discussion, see my 1998 book *In Defense of Pure Reason*, pp. 174–80.

that I successfully use—and that the same is true of my language of thought. On this view, the intuition that I understand what I am thinking about—or what I am talking about—in any stronger sense, one that is not merely disquotational, is merely an illusion. But here again we have a view that it seems to me would appeal to no one who was not motivated by the conviction that materialism must be true.

My conclusion is that the language of thought view has nothing useful to say about the most obvious sort of intentional content: the intentional content that is involved in having something explicitly and consciously in mind. Nor do I know of any other materialist account that does any better in this regard. There is perhaps room for dispute about just how important conscious intentional content is in relation to the causation and explanation of behavior, but no plausible way to deny that it genuinely exists. Thus with respect to intentional content, as with the case of qualitative content, materialism seems to be utterly bankrupt as a general account of mental states and to be held merely as an article of faith.

5. WHAT IS THE ALTERNATIVE?

The last two sections serve merely to strengthen and deepen the fundamental objection to materialism already offered in section 2: consciousness genuinely exists; materialism can offer no account that explains consciousness (or of the specific varieties thereof) or shows it to be merely material in character; therefore (at least in the absence of any strong antecedent argument or presumption in favor of materialism), the indicated conclusion is that materialism is false. There is more in heaven and earth than is dreamt of in materialist philosophy.

But what do I mean by more? Here, as I see it, there is very little that can be said in our current state of knowledge, so that the main result is that we have very little understanding of consciousness—or, given the arguably central role of consciousness, of mentality in general.

In the first place, there is no clear way in which the objections that I have raised against materialism support the classical substance dualist position. Positing a separate mental substance that is characterized in almost entirely negative terms does nothing very obvious to explain consciousness in general, or qualitative and intentional content in particular. As far as I can see, the main appeal of substance dualism is that the account of the supposed mental or spiritual substances is far too vague and sketchy to provide the basis for any very clear argument that such substances could *not* be the locus of consciousness. But this negative point hardly counts as an argument in favor of such a view.

The obvious alternative is 'property dualism': the view that human persons and perhaps other kinds of animals have non-material or non-physical properties in addition to their physical ones, with at least the main such properties being

the various kinds of consciousness, including the central ones that have been discussed here. In a way, this view seems obviously correct. The properties in question genuinely exist and seem, on the basis of the failure of materialism to explain or account for them, to be clearly non-material in character. But without some further explanation of what such properties amount to or of how they could be properties of a mostly material organism—or, for that matter, of an immaterial substance—the property dualist view yields little in the way of real understanding and hardly counts as a serious account of the nature of mental states.

One somewhat more definite result can, I think, be derived from the discussion of conscious intentional thought. If when I think consciously about things, I am able to know what it is that I am thinking about without knowing anything further about external relations, then what the states in question are about must apparently be an *intrinsic* feature of them: they must have *intrinsic intentionality*, as opposed to an intentionality like that of language (including a language of thought) that is derived from external relations. When I am consciously thinking about, say, trees, there must be something about the intrinsic character of my state of mind that makes it about trees (and in a way that is immediately apparent to me). Here we have a conclusion that very few would accept and that many would regard as virtually absurd. All I can say is that it seems to me clearly required by the facts of the situation.

But how could the intrinsic character of a state definitively pick out something external to it in this way? I do not claim to have anything like a clear answer to this question, but I will indulge in a bit of what seems to me initially plausible speculation. First, I offer the surmise that what is needed to account for intrinsic intentionality in general is an account of two sorts of intrinsically intentional elements: first, intrinsic reference to *properties* of various kinds; and, second, intrinsically *indexical* content.

About the latter of these, it is reasonably plausible to suppose that indexical content of all kinds can be reduced to an indexical reference to the self, with other things, including other places and times, being indexically specified by appeal to their relations to the self. Such a view has sometimes been suggested by others as well,[22] but I have no space to develop it further here.

Intrinsic reference to properties seems more difficult. Including anything in a state that merely in some way *stands for* or *represents* a property does not seem to yield intrinsic intentionality, since the reference to the target property will also depend on the external relation between this representing element and that property itself. Having a symbolic element that stands for the target property in question obviously will not work, for reasons that we have already seen in the earlier discussion. But having a representing element that *resembles* the target property also seems inadequate. If the representing element resembles the target

[22] See, for example, David Lewis (1979a).

property by having some other distinct property, then the connection to the target property seems to depend on the relation of resemblance in a way that makes the reference to the target property no longer intrinsic. It is also hard to see how someone who has direct access only to the resembling property would be able to be aware that they were thinking of the target property (or of something else that was picked out by appeal to the target property). Such a person would seemingly have only the resembling property and not the target property explicitly in mind.

Thus what seems to be required is that the intrinsically intentional state actually involve, in some way, the target property itself. Nothing else seems adequate to make the reference to that property both intrinsic and in principle accessible to the person having the thought. Obviously though this cannot in general involve the intrinsically intentional state or some component of it literally instantiating the target property, for obviously we can think about lots of properties that are not literally instantiated in our intentional states. Elsewhere I have speculated that what might be involved is the state or some component instantiating a *complex universal* that has the target property as an ingredient in some appropriate way.[23] But while this proposal seems to have in a way the right sort of structure, I do not really claim to have even an initial understanding of what it would involve or how it would work.

My conclusion remains almost entirely negative. We can see that consciousness exists, and we can see what this specific sort of consciousness in particular would have to involve—namely intrinsic intentionality. And seeing what intrinsic intentionality in turn would require makes it, if anything, even clearer that there is no reason at all to think that a merely material state could have this characteristic. But how consciousness in general or intrinsic intentionality in particular can be explained and accounted for is something about which, if I am right, we know almost nothing.

[23] See BonJour (1998: 180–6).