4 Qualia

If Descartes is right, pinching yourself will not suffice to prove that you are awake. It may suffice, however, to prove something more philosophically momentous: that materialism is false. That, at any rate, is the claim of a number of recent anti-materialist arguments in the philosophy of mind. The feel of the pinch - the subjective, "inner" element that makes it true that there is "something it is like" to be pinched - appears to be distirict from and additional to objective "outer" phenomena such as the reddening of the pinched skin, the stimulation of nerve endings, or indeed anything material or physical. It seems, in short, to be immaterial or nor-physical, and, if it is, its very existence refutes the materialist claim that everything real is really material.

Qualia -'the feel of a pinch or an itch or a pain, the taste of apple or whiskey, the redness of a fire engine or an after-image, and so on for all the sensory modalities - constitute, in the minds of many philosophers, the most serious challenge to materialism. The little said about them so far in this book has perhaps given an intuitive sense of why this is. And then again, perhaps not; for it is easy to understand why someone might not be clear on exactly what the problem is. After all, isn't the pain of your toothache, in an obvious sense, in your tooth? And if it is, doesn't that show that it is physical? Your tooth is physical, after all, so wouldn't anything in it - blood vessels or pain - have to be physical too? But the pain isn't "in" your tooth in quite the same sense

in which blood vessels are - you can't observe or pin-point the pain the way you can the blood vessels - and that should be a hint that there might indeed be something philosophically mysterious going on here. In any case, a number of recent arguments have attempted to make plain precisely what qualia are and how they are supposed to be impossible to account for in purely material terms.

The lnverted spectrum

The idea of the "inverted spectrum" has a long history in philosophy, going back at least to Locke, but it has served recent philosophers well

as a means of motivating the problem of qualia. It goes like this: it seems possible that another person, even one who is physically, behaviorally, and functionally identical to you could have color experiences which are inverted relative to your own; that is, what you see when you look at what you both call red, for instance, is what the other person sees when he or she looks at what you both call green, and vice versa, and this difference would, nevertheless, not register in what either of you said about red and green objects or in how you interacted with them. If you were somehow able to look inside the other person's mind when he or she was looking at what you both call red, you would say "Wait a second, that's what I would call green!" and if he or she could look inside your mind when you're looking at what you both call green, the other person would say "Wait a second, that's what / would call red!" Since neither of you can do this, however, the difference in the subjective character of your experiences goes unnoticed. The scenario is similar to the difference in experiences between those who are color-blind and those with normal vision: color-blind people can make many of the same discriminations between objects that everyone else can, so their color-blindness can, in principle, go undetected for quite some time. From the "outside" it might appear that the experiences of color-blind people and those of naturally sighted people are identical, but they are not. The inverted spectrum scenario is just an extension of this, a case where the difference between your experiences and those of the other person is absolutely undetectable from the outside. It would seem to follow from the possibility of such inverted color experiences that facts about color qualia - about what it is like to experience red and green - are facts over and above the facts about one's physical make-up and functional organization; for those latter, purely physical facts would, in this case, not be enough by themselves to determine the nature of the color experiences one is having. But then materialism, which holds that the physical facts involved in color experiences are all the facts there are, would seem to be false.

Similar scenarios can be described in which what is inverted are not color, but some other kind of qualia. we can imagine, for example, that what you taste when eating what you and other people both call sweet is what they taste when eating what you would both call savory; that what you feel when experiencing what you would both call pain is what they feel when experiencing what you would both call pleasure, and so on.

The color inversion is probably the easiest to imagine because of its similarity to the real-world phenomenon of color brindness. But it also suggests how the materialist might be able to get around the problem. The inverted spectrum scenario will only be a difficulty for materialism if indeed there is absolutely no way in principle for the inversion to be detected from the outside - no way for it to manifest itself in differences in behavior, or in differences in the functional organization of you and the other person. But there seems to be good reason to doubt that this would be impossible in principle. As philosophers of mind like C. L. Hardin and Austen Clark have emphasized, the scientific study of color and color vision has revealed there to be highly complex relations between the various colors, such that any particular color can be given a detailed description in terms of its relations to the others. These form, when made thoroughly explicit, an abstract structure sometimes referred to as "color space," a system of relations within which each color can be given a precise location. This structure appears to be asymmetrical, however. Features characteristic of one part of color space - ~~the "warmth" of red,~~ say - are absent in other parts, such as the area where blue lies, which is characteristically"cool." The number of shades that can be discriminated in the case of one color might not match the number discriminable in the case of another: we believe, for instance, we can discriminate more shades of red than of yellow. And so forth. But these asymmetries would surely manifest themselves in the functional organization and behavior of color perceivers whose color qualia had been inverted: if you saw what I would call blue whenever you looked at what we both call red objects, you presumably would not, as I would, react to those objects in a way that corresponded to the "warmth" that their color seems to me to exhibit; if you saw what I would call yellow when you looked at what we both call red objects, you surely would not be able to distinguish the same number of shades of their color as I would be able to; and so forth. It seems likely, then, that a qualia inversion would in principle be detectable "from the outside" - from differences in the physical-facts.

Materialism, which holds that the physical facts are all the facts there are, would thus not be refuted by the inverted spectrum idea after all.

It is sometimes replied to this that even if our color experiences could not be inverted undetectably, it is nevertheless possible that there could be other creatures who perceived two different colors whose relations were symmetrical, so that an inversion of their experiences would be undetectable from the outside. If so, then the facts about their color experiences would be facts over and above the facts about their physical constitutions and functional organization, and the anti-materialist implications of the inverted spectrum idea would still stand. But it is not at all clear that this is possible. What exactly would these hypothetical colors be like? Certainly not like our colors (e.g. red and blue), whose structure is asymmetrical (e.g. warmth versus coolness). What we need to conceive of, then, in order to be sure that the suggestion really is possible, are colors totally unlike ours, whose structure is sfmmetrical and yet could be inverted without detection. But it is hard to see how anyone could, with any confidence, claim that this really is conceivable. In particular, it is hard to see how we can be confident that two colors whose relations were entirely symmetrical would count as different colors in the first place. The inverted spectrum scenario thus seems difficult to salvage as a decisive argument against materialism.

The “Chinese nation” argument

Even the inverted spectrum scenario doesn't claim to show that the physical features of the nervous system, behavior, etc. are completely divorced from qualia. What is at issue is whether the purely physical properties of your nervous system are sufficient to determine the precise character of your qualia; that you have qualia of some sort or other is not in question. But there is another famous thought erperiment that attempts to show that at least the functionalist version of materialism - the version which, as we've seen, is currently the most popular - fails to explain not only the specific character of qualia but even why we have any qualia at all. This is the "Chinese nation" argument, named after a thought experiment devised by Ned Block.

Functionalism, as we've seen, takes mental states to be properly definable in terms of their causal relations, not in terms of the particular kind of stuff in which those causal relations happen to be instantiated. A belief is a belief, whether it is realized in the firing of neurons or in the

passing of electrical current through computer circuitry. Anything that plays the requisite functional role will do the job. If computer chips can perform the same function as neurons which is, basically, nothing more than the receiving and transmitting of simple signals - then they can, when organized into a system as complex as the system constituted by our neurons, generate a mental life just as rich as ours. But what is true of computer chips should, if functionalism is correct, be true of any number of other possible elements. We can imagine, for instance, that an enormous number of people - the population of China, let's suppose - could be mobilized to interact with one another in a way that exactly parallels the interaction of neurons in the brain. At the most basic level, those neurons merely send signals to fire or refrain from firing to other neurons. So we can imagine that each member of the population is given instructions to do something similar, perhaps by sending signals to each other via walkie-talkie or cell phone to the effect that the people receiving them should either go on to send a further signal down the line or to refrain from sending one. Suppose ~~also~~ that this vast network of people is connected, via a radio transmitter, to a complex robotic body sophisticated enough in its construction to receive, through its artificial sensory organs, just the sorts of information our senses receive and to exhibit just the sorts of behavior we exhibit. The network of walkie-talkie or cell phone-wielding signalers serves, collectively, as the "brain" of this robotic body. When the robot is kicked in the shins, the artificial nerve endings in the legs send signals up to the radio transmitter in the robot's head which in turn sends signals to a few hundred thousand members of the walkie-talkie network, who in turn send signals to a few hundred thousand others, who in turn send signals to others, and so on until at the end of the line the last members of the network send signals back to the robot, as a result of which the robot yelps"Ouch!" and rubs its shins. The signals sent between the members of the network parallel exactly the signals sent between neurons when a human being is kicked in the shins, and produce the same behavioral response. And we can imagine that the network of Chinese signalers is so organized that their interactions parallel those of neurons in every other respect as well, so that in general, the robot body behaves exactly as we do in exactly the same sorts of circumstances: conversing with others, laughing at jokes, and crying at injuries.

As in the case of the original robot example we used to motivate functionalism, we have, in this robot controlled by the population of China - "China-head," as some philosophers have affectionately dubbed it - a system which is functionally identical to us: it produces the same sorts of behavior in response to the same sorts of stimulation, and via exactly parallel intermediate processing, but instantiated in walkie-talkie-using people rather than in neurons. If the functionalist is right, this system, however eccentric' should have mental states just like ours, and in particular qualia just like ours. But would it? It is, for instance, hard to believe that when you kick China-head in the shins, the entire population of China collectively, as a vast super-mind, feels pain! But if it doesn't, then functionalism is false: for if a system could be functionally identical to us and yet lack qualia, then there is more to having a mind, and in particular more to having qualia, than having a certain sort of functional organization.

That is the conclusion Block and others take to be the intuitive one. But the "Chinese nation" argument, like the inverted spectrum argument, seems less than conclusive as a qualia-based argument against materialism. For it seems that the gradual transformation scenario which, is we saw in the previous chapter, the functionalist can use to defend the claim that a Data-type robot would be conscious, can be adapted for use against the "China-head' example. Consider a case we can call the "spaghetti-head" scenario.

Even if you doubt that China-head would be conscious, you surely have no doubt that you are. Now imagine that you are kidnapped by mad, philosophically inclined neuroscientists who strap you to a table in'their laboratory and remove the top of your skull, exposing your brain. Suppose they've figured out how to disentangle the billions of tiny nerve fibers constituting it in a way that their functioning is not affected. Slowly and carefully, they hang them from hooks above the table, labeling each one with a number. Then they treat them with a special chemical that allows the fibers to be stretched almost indefinitely without breaking or losing their conductivity. Eventually the room becomes filled with billions of tiny strands hanging from the ceiling. All this time, though, you continue to be as capable of having thoughts and experiences as you were before, and notice no difference in your mental life. Of course, all

of this is science-fiction of the sort not likely ever to be realizable. But it seems perfectly conceivable, and thus metaphysically possible.

Now suppose that, as in the gradual transformation described in chapter 3, each of your stretched-out: neurons is gradually replaced only this time, they are not replaced with computer chips, but with people. Specifically, when a neuron is removed, the neuroscientists attach a radio unit to each neuron with which it had been connected, .and give another radio unit to the person replacing it. Instead of sending an electro-chemical signal, the neurons which previously triggered the replaced neuron now send a radio signal which is picked up on the human replacement's radio, and that person in turn sends further radio signals, in lieu of electro-chemical ones, to other neurons, just as the original neuron used to. Suppose that at first only a hundred or so neurons are replaced in this way. As in our original replacement scenario last chapter, it seems highly implausible that this would affect mental functioning in anyway: the people with the radio units are doing exactly what the original neurons did, so your mental life - including your qualia - should be just as they were before.

The reader has no doubt guessed where all this is going. We can imagine that all your neurons are eventually replaced in this way perhaps by the population of China. Spaghetti-head is transformed into China-head. Yet at no point in this gradual transformation is it plausible that your qualia disappear, for as in the computer-chip replacement scenario described in chapter 3, the functioning of your nervous system remains exactly the same, whether composed of neurons or people with radios: why, then, should it cease generating the mental states it did before? At the very least, it seems possible, given the gradualness of the change, that your qualia would remain the same. But then Block's original "Chinese nation" example seems much less compelling. If you, having been gradually transformed into China head, would remain conscious, why couldn't the original China-head - who is, after all, functionally identical to you - also be conscious? It seems at least arguable that it would be: in which case Block's argument also fails decisively to refute functionalism.

The zombie argument

For all that has been said so fan it might still seem that there is something fishy about the suggestion that China-head would truly be conscious. In any event, many critics of materialism hold that the basic thrust of the Chinese nation argument - that it is metaphysically possible for a creature functionally identical to us nevertheless to lack qualia - can be defended without having to appeal to systems as eccentric as the one Block envisages. This brings us to the "zombie argument."

It seems perfectly conceivable, and thus metaphysically possible, for there to be a creature which is (unlike China-head) physically identical to you, down to the last molecule - one which looks and acts exactly the same, which is absolutely indistinguishable in its material and functional characteristics even after the most detailed examination – and yet is totally devoid of conscious experience. When you step on a tack, there is damage to the skin of your foot, stimulation of the nerve endings, signals sent up the leg to the spinal cord, a consequent reflexive pulling away of your foot, further signals sent up to the brain, and complex neural processing that climaxes in you clenching your teeth and yelling "Ouch! “Also, associated with all this physical activity, there is a subjective throbbing feeling of the sort we normally associate with pain. When the creature steps on a tack, there is also damage to the skin of its foot, stimulation of its nerve endings, signals sent up its leg to the spinal cord, a consequent reflexive pulling away of its foot, further signals sent up to its brain, and complex neural processing that climaxes in it clenching its teeth and yelling "Ouch!" But there is in this case no subjective feeling of pain, or any other conscious experience associated with these physical processes at all. Anyone observing the creature from the outside would be unable to tell it apart from you, for your physical characteristics and behavior are identical. Indeed, just like you, the creature would, if asked whether it was conscious and whether it was really in pain, respond, with apparent indignation, "Of course I am!" Still, there is a dramatic difference on the inside: in your case, there is a rich and vivid stream of sensations and experiences; in its case, all is dark. Such a creature is what philosophers of mind have come to call a zombie: a creature exactly like us in all its behavioral, physical, and functional properties but totally lacking qualia.

If zombies are metaphysically possible, then materialism would seem to be false, for it holds that behavioral, physical, and functional

properties are all the properties there are, and that they are entirely sufficient for the having of any mental state. But the possibility of zombies entails that facts about qualia are additional to, over and above, the having of behavioral, physical, and functional Properties: if a creature could have all those properties and yet lack qualia, then to have mental states involving qualia is something more than just having those properties. The zombie argument is the flip side of the conceivability argument for dualism discussed in chapter 2. There the claim was that it is conceivable, and thus metaphysically possible, for the mind to exist apart from the body, brain, or any physical substrate at all. Here the claim is that it is conceivable, and thus metaphysically possible, for a fully functioning body and brain to exist without any mind presnt at all (or at least without certain aspects of the mind - qualia - being present). The upshot is the same in both cases: the mind is not merely the body or brain (or anything physical for that matter), but is something additional to them.

The zombie argument also sometimes goes by the name of the conceivability argument, though unlike the argument of chapter 2, it attempts to undermine materialism without necessarily committing itself to full-blooded Cartesian substance dualism. One could accept the zombie argument without holding that the mind can exist entirely apart from the brain and body; the claim would just be that even if conscious experiences are causally dependent on the brain for their existence, they are nevertheless not reducible to (or metaphysically supervenient upon) purely physical or functional properties of the brain. So some of the objections the materialist might make against Descartes's brand of dualism (to the effect that the mind seems too dependent on specific features of the brain to exist completely independently of it) are without force against this argument. The argument is also sometimes called the modal argumenr against materialism because, like the argument of chapter 2, it appeals to such modal notions as metaphysical possibility; indeed, an early version of this argument was presented by Kripke, whose work on possibility and necessity has been enormously influential in contemporary philosophy of mind, as our earlier discussion of the conceivability argument indicated. And the defense of that argument made in chapter 2 by appealing to some of Kripke's ideas would also apply more or less without alteration to defending the zombie argument

against any parallel objections one might think to raise (such parallel objections being, indeed, the standard objections to the zombie argurnent).

The zombie argument thus seems to exacerbate the problem for materialism posed by the original conceivability argument: it is at least as strong as the latter, and maybe stronger, since it shows that the critique of materialism by no means stands or falls with the acceptability of substance dualism.

The knowledge argument

The zombie argument tries to show that physical reality does not, on its own, add up to mental reality. A related argument, which reinforces this basic idea, tries to show that knowledge of physical reality does not on its own add up to knowledge of mental reality. It is accordingly generally known as the knowledge argument, and derives from the contemporary philosopher Frank Jackson.

Jackson asks us to consider Mary a neuroscientist living in the far future when we have a complete knowledge of the details of the structure and functioning of the nervous system. Mary is in the unique situation of having lived her entire life in a black-and-white room, interacting with the outside world via a black-and-white television monitor. So she has never had any experience of color. (We can even imagine that she has always worn a suit that covers her entire body, and which has kept her from seeing the color of her skin and hair, etc.) While in this room she has come to master the science of the brain, and in particular she has acquired a thorough knowledge of the physics and physiology of color perception. She has never seen the color red herself, but she knows exactly what happens in the eyes, nervous system, and on the surface of the object whenever anyone does see red. She knows down to the last detail, that is to say, all the physical facts there are to know about the perception of color. Now let's imagine that one day Mary is allowed to leave the room, and upon her release she is shown a red apple in full living color for the very first time. Will she learn anything from this experience? Surely she will: she will learn what it is like to see red. And what this shows, according to the argument, is that materialism is false.

The reasoning is this. Materialism claims that the physical facts about perception and the like are all the facts there are. But Mary, hypothetically, knew all the physical facts there were to know about perception - the sorts of facts that could be written down in neuroscience textbooks or conveyed in lectures heard over the television monitor. Yet she did not know all the facts there were to know about perception, because she learned something new about it upon leaving the room - and you can't learn something you knew already. So what she learned must be a non-physical fact. In particular, knowledge about qualia - about what it's like to see red, for instance - must be knowledge about something non-physical.

The suggestion that knowledge of all the relevant physical facts cannot yield knowledge of all the facts about conscious experience has also been illustrated vividly in an example given by Thomas Nagel. Bats, Nagel notes, navigate via senses very different from our own: where we rely chiefly on vision and hearing, they use a kind of sonar or echolocation, putting together a sensory map of the external world by emitting shrieks and then registering the sound waves that bounce backto them from the objects in their immediate environment. The experiences bats have in perceiving the world in this way must be radically dissimilar to ours. Scientific investigation into the structure and functioning of a bat's nervous system maywell give us insight into the mechanics underlying its perceptions. But the nature of the perceptual experiences themselves - what it is like to be a bat - cannot be revealed by such inquiry Nagel argues. For science gives us only the objective, third person facts about any phenomenon, leaving aside any aspect tied to a particular point of view. But it is only from the particular, subjective point of view of a bat that a bat's experiences can be understood. Materialistic scientific accounts must necessarily be inadequate to capture all the facts about a bat's consciousness - or any consciousness, for that matter.

One response sometimes made to arguments like this is that they simply assume that future neuroscience won't be able to explain all there is to explain about conscious experiences: how can we know for sure that Mary wouldn't know what it is like to see red, simply from having mastered the material in her textbooks while in the black-and- white room? There are two problems with this suggestion. The first is that it seems intuitively implausible. Any facts the neuroscientists of the future

are likely to discover are bound to be facts of the same general sort they already know: facts about how neurons are wired, or about which biochemical substances are involved in which processes. It is hard to see how any further knowledge of that sort - of yet more objective, third-person phenomena - could reveal the subjective, first person facts about what it is like to experience red or to get about by echolocation; there is just a basic and straightforward conceptual difference between the former sort of fact and the latter. The second problem is that the suggestion at hand seems inevitably beset by the same indeterminacy that plagues some versions of physicalism, as we saw in the previous chapter: what if the way neuroscientists of the future explain conscious experience is by positing non-physical properties? This would vindicate the knowledge argument rather than undermine it. Yet there is nothing about the current course of neuro science that can reasonably lead us to expect any other way in which it might explain consciousness.

More formidable responses to the knowledge argument usually proceed by conceding that there is a sense in which Mary would learn something upon leaving the room, even though she's mastered the neuroscience of the future. The strategy is then to argue that what she learns can, when right), understood, be seen not genuinely to threaten materialism. Paul Churchland argues that on leaving the room, Mary would not actually learn any new facts; rather, she would just learn, in a new way, facts she already knew. So since she already knew all the physical facts, and there are no new facts (non-physical or otherwise) she learns after leaving the room, the conclusion that the physical facts cannot be all the facts there are is blocked. Churchland elaborates upon this suggestion by appealing to Russell's famous distinction between 'knowledge by acquaintance" and "knowledge by description": you might now know about giraffes only by descriptions you've heard or read in a book, but you might someday know about them by becoming directly acquainted with them in perceptual experience; similarly, Mary, while still in the room, knew all the facts about the experience of red only by description, and then becomes directly "acquainted with those very same facts after leaving the room.

One possible objection to this argument is that it seems implausible to suggest that Mary doesn't learn a new fact on leaving the room: surely the fact that red looks like this (where "this" refers to the

immediate sensation she has of the color) is a fact she did not know before leaving the room, but learns afterward. Another problem is that the Russellian distinction churchland appeals to is not as philosophically neutral as it might appear. Russell himself held that all we really know by acquaintance are, not external physical objects like giraffes, but rather (what philosophers these days would call) the subjective qualia we normally supposed to have been produced by such external objects; the external physical world in its totality is something we know only indirectly, by description. This goes hand in hand with the sort of indirect realist theory of perception discussed in chapter l, of which Russell was a proponent (as is laclson, for that matter). It also raises the question of precisely what these qualia are with which we are directly acquainted; Jackson and (as we'll see in the next chapter) Russell take them to be irreducible to the sorts of properties revealed by physical science, properties which, unlike qualia, we cannot know by acquaintance. So to appeal to Russell's conception of knowledge by acquaintance can hardly help Churchland in rebutting an argument against materialism. But to reject Russell's conception and insist instead that knowledge by acquaintance does not involve knowledge of non-physical qualia would be to beg the question. Either way it seems that Churchland's response to Jackson's argument fails.

Another response is put forward by David Lewis, who, like Churchland, denies that what Mary learns is a fact she didn't know before. Rather, the knowledge she gets is knowledge of new abilities: knowledge of how to do something rather than knowledge that something is the case, and in particular knowledge of how to recognize red objects, the ability to imagine red, and so forth. But this reply-seems to have problems parallel to those undermining Churchland's: for one thing, it seems implausible to assert that Mary learns no new facts, since knowledge that red looks like this (referring to a subjective sensation) is knowledge of a new fact; for another, the distinction Lewis appeals to is itself not necessarily a neutral one. Mary may well gain new abilities or knowledge upon leaving the room, but it is arguable that some of those abilities are gained only because she learns new facts: Mary now has the ability to imagine what red looks like, but only because she has also learned the fact that red looks like this.

Robert van Gulick presents a somewhat technical reply to Jackson's argument. He claims that what Mary gains is knowledge of a new concept, and that if she also learns new Propositions this is so only on a fine-grained scheme of individuating or distinguishing between propositions. What this means can best be explained by example. Whether the proposition that water freezes at 32 degrees Fahrenheit and the proposition that H_2O freezes at 32 degrees Fahrenheit are the same proposition depends on whether we individuate propositions in a fine- or coarse-grained mode. A fine-grained mode would be one which took account of the fact that "water" and "H_2O" are associated with different concepts (even though they refer to the same substance) and thus would count these propositions as distinct; a coarse-grained mode would ignore the difference in concepts and (since "water" and "H_2O" refer to the same substance) count them as identical. Similarly the proposition that 5 + 7 = 12 and the proposition that 38 is the square root of 1,444 are the same proposition on a coarse-grained mode of individuating propositions (one that takes account only of the fact that these mathematical propositions, being necessarily true, both have exactly the same truth value in every possible world); but they are different propositions on a fine-grained scheme, one that takes account of the different concepts associated with "5," " +," "7,"'12,""38 "square root," and " 1,444." In the first example, it is clear that even if we count the propositions as different, the fact they refer to is the same: water is identical to H_2O, so the fact that water freezes at 32 degrees Fahrenheit is the same fact as the fact that H_2O freezes at 32 degrees Fahrenheit. Similarly, van Gulick suggests, even if Mary, having learned a new concept after leaving the room, is thereby also able to learn a new proposition, it would not follow that the fact that proposition describes is a fact she didn't already know. Perhaps it is a physical fact of the same sort she already knew while still in the room.

As with the other responses to the knowledge argument, one could object to this one that it seems intuitively implausible: the fact that red looks like this (where "this refers to an immediate sensation) seems obviously to be a different fact than the fact that Mary is in a brain state of type B (or whatever). Of course, van Gulick might suggest that the way things seem might nevertheless in this case be wrong: it might also seem to someone ignorant of chemistry that the fact that water freezes at 32 degrees Fahrenheit is a different fact from the fact that H_2O freezes at

32 degrees Fahrenheit, even though they are in reality the same. But it isn't clear that this suggestion will work. After all, few people would find it a satisfactory defense of the highly dubious claim that the fact that 5 + 7 = 12 is the same fact as the fact that 38 is the square root of 1,444. In the case of this mathematical example, we surely have two different facts, not just two different fine-grained propositions. Indeed, it is partly our sense that this is so that leads us to see the need for a fine-grained mode of individuating propositions in the first place: we don't suppose this is necessary merely in order to take account of differences in concepts, but also because the propositions of which concepts are constituents often seem (as in the mathematical example) to be about different facts. But the suggestion that the facts that Mary learns on leaving the room are the very same facts as those she knew before seems just as intuitively implausible as the suggestion that the mathematical facts in our example are the same. And if such an implausibility is, in the one case, itself precisely what leads us to accept a more fine-grained account of mathematical propositions - so that it would be absurd to suppose that one could defend the claim that the mathematical facts in question are the same by appealing to a fine-grained account - then it would be (equally) absurd and implausible to suppose that one could refute the knowledge argument by a parallel appeal to a fine-grained scheme of individuating propositions. In other words, it is in part precisely because it seems so intuitively plausible that facts about qualia and physical facts are just different sorts of fact that we find a fine-grained mode of individuating propositions about them to be necessary in the first place. So it won't do to appeal to such a mode in order to defend the claim that they aren't different.

Subjectivity

Most of the criticisms of the knowledge argument are more or less along the same lines, and would therefore be open to similar objections. But there is another possible reply, suggested by what was said earlier about the inverted spectrum scenario, which may be more formidable. Suppose that each color can indeed be given a precise location in color space, and thus analyzed in terms of its relations to every other color. It then seems possible, at least in principle, that on might be able to deduce the nature

of one color from its relations to the others. Consider a simple example involving three very close shades of blue, A, B, and C, where A is the lightest, C the darkest, and B intermediate. It is certainly plausible that someone who had only ever experienced A and C would be able to figure out what it would be like to experience B simply by considering its relations to A and C (the relations being "darker than" and "lighter than"). By extension, it may also be plausible to suggest that someone who had never seen orange could, in principle, determine what it would be like to experience it if he or she had exrperienced red and yellow: one could deduce the appearance of orange from its being similar to, and intermediate between, these other colors. Why not conclude, then, that someone who had had at least some visual experience - of black and white, of gray as intermediate between them, of light and dark - might in principle be capable of deducing what the various colors looked like based on a sufficiently detailed description of their relations? Why not conclude in particular that Mary- who studied the theory of color and the structure of color space - would have been able in principle to deduce what it would be like to experience red while still in the room, so that she would in fact not have learned anything new when leaving it?

This sort of strategy could in theory be extended to all qualia - auditory, tactile, olfactory and gustatory as well as visual - which could all be described in terms of their relations to other qualia of the same sort, and even their relations to qualia of different sorts: "warmth," "coolness," "hardness" softness" "sharpness" smoothness' seem to be qualities applicable to many different kinds of qualia, so that (to some extent at least) visual qualia can be described in terms of their similarity relations to auditory qualia, auditory qualia in terms of their similarity relations to tactile qualia, and so forth. Rudolf Carnap (189l-1970) attempted just such a detailed and systematic analysis of all qualia in terms of their relations to each other, which relations he took to be grounded ultimately in the basic relation of "recollection of similarity." If such an analysis could be carried out completely, then it is arguable that anyone thoroughly familiar with it could, on the basis of even the most limited sensory experience, determine what it would be like to have any experience that he or she has never in fact had.

This approach seems promising, though it would take a great deal of argument convincingly to defend it. But even if successful, the critic

of materialism could hold that this strategy would not undermine the deeper truth captured by the knowledge argument, in Nagel's version more than in Jackson's. That truth is, arguably, just

this: while Mary might at least in principle be able to deduce, from what she knows while still in the room, what it is like to experience red, she would not be able to deduce from it why it is like anything at all.The real mystery is not that red "feels" specifically like this rather than that it is that it has any "feel" in the first place. Nagel captures the problem by noting that it is the fact that there is "something it is like" to be conscious that makes consciousness so difficult to account for in purely material terms. The zombie argument captures it by suggesting that it is metaphysically possible for there to be creatures physically identical to us but without consciousness, creatures who exhibit exactly the same behavior - and thus, for example, make exactly the same discriminations between red and other colors – but who do not experience red, for whom there is nothing it is like to discriminate red from other colors. That there is something it is like for us to experience it would seem to be a further fact about us, over and above the physical ones.

This goes hand in hand with Nagel's point that a conscious being is one with a first-person point of view on the world, who is a locus of subjectivity. Consciousness of what an experience is like is always consciousness of what it is like "for me,' for a subject of experience; and for Mary to deduce what experiencing red would be like from its similarity relations to other experiences presupposes that she is a conscious subject for whom it would be similar. One might think to deflate this notion of subjectivity by suggesting that lots of purely physical things have points of view on the world as well - a camera, for instance, which can photograph only what is in front of it; its images produced by reflecting its particular point of view - so that it shouldn't be so mysterious why we, with our specific sensory organs and physical limitations, should have points of view too. But such a suggestion would seem fallacious. A camera is just a mechanism sensitive to light such that it can be used to generate patterns on film that correspond to the light patterns reflected by physical objects. It has no literal "point of view," for

it doesn't view anything in the first place in the sense in which we do. It is we who understand the pictures the camera produces to have significance - indeed, it is we who regard them as pictures rather than splotches of chemicals on paper. It is also true that the particular point of view any of us occupies is, like the camera, limited by our specific position in space and the physical constraints imposed by the structure of the human body. But (to make a point that parallels the point made above about the experience of seeing red) it is not our having this or that particular point of view that is claimed to be difficult or impossible to explain in materialistic terms; it is rather our having any point of view at all that is mysterious.

ln the dualist's view, that science, at least as understood by materialists, cannot in principle solve this mystery seems to follow necessarily from the very nature of scientific explanation: it is not a matter of our not yet having gathered all the relevant neurological evidence or hit upon the right theory. For, as noted in the last chapter, the method of modern scientific explanation has historically been precisely to carve off and ignore the subjective, observer-relative aspect of any phenomenon it investigates and identify such phenomena exclusively with the objective, third-person residue which remains. We can take the explanation of temperature as a paradigm. A hoary philosophical example illustrates the subjectivity of temperature considered as a felt experience: someone who first puts his or her right hand in a bucket of ice cold water and his or her left in a bucket of hot, then puts both in a bucket of lukewarm water, will find that the lukewarm water feels warm to the right hand and cold to the left. We can also imagine extra-terrestrials who would feel what we would call coolness when putting their hands (or tentacles) in hot water and heat when putting them in ice cold water. If by "heat" and "cold" we mean the subjective sensations or feelings produced by hot and cold objects, there is no objective fact about whether a particular object is hot or cold. Science thus ignores subjective feelings and instead defines (or re-defines) heat and cold exclusively in terms of the objective, mind-independent physical facts which (in us, anyway) cause the relevant sensations: facts about mean molecular kinetic energy. But if the method of science is in every case to strip away the subjective appearance a phenomenon exhibits and, as it were, push it into the mind, it seems obvious that the same procedure cannot in

principle be applied to an explanation of the mind itsetf: for the mind just is (in part) the collection of the subjective appearances of the things it experiences; the subjective element cannot in this case be stripped away without thereby stripping away and ignoring the very phenomenon to be explained - in which case it hasn't really been explained at all.

Subjectivity - comprising the phenomena of being Present to an experiencing subject, of being directly accessible only from the point of view of that subject, and of being capable of existing in experience even when (as in dreams or hallucinations) an apparent objective correlate of the experience does not exist - thus appears to be the essential core to the concept of qualia, and the feature that is most plausibly inexplicable in physical terms. Philosophers often attribute other supposedly problematic features to qualia, such as ineffability and intrinsicality, but to a very great extent these appear to be reducible to or parasitic upon subjectivity. For example, qualia seem ineffable only because our language is typically used to communicate thoughts about objective, public phenomena, and words are typically learned by reference to such phenomena; communicating thoughts about private and subjective phenomena thus seems difficult or impossible. To the extent that qualia are ineffable, this is just a consequence of their being subjective.

Qualia are often claimed to be intrinsic in the sense of not being analyzable in terms of their relations to other things, for example; in terms of the causal relations functionalism claims all mental phenomena can be analyzed in terms of; for, as was suggested by the zombie argument, it seems logically possible for any such set of causal relations to exist without qualia. But here too subjectivity seems to be what's really at issue. It is because qualia are not analyzable into relations instantiated in objective, third-person phenomena - causal relations between firing patterns in clumps of neurons, say - that they seem to be intrinsic. Yet this leaves open that they may be analyzable into subjective, first-person similarity relations of the sort Carnap, Clark, and Hardin have tried to elucidate: that they may well in this sense be both irreducibly subjective and yet non-intrinsic. Indeed, it is arguable that it is precisely because they are so analyzable that we can communicate about them despite their subjectivity (so that they are not ineffable in the strict sense): if we were not able to describe and convey to one another the systematic similarities and differences between qualia, we would not be able to know (as we

surely do know) that we are all talking about the same phenomena when we discuss qualia and argue about whether materialism can account for them. Our knowledge of the relational structure of qualia makes our claims about them cognitively meaningful and rationally assessable, despite the fact that the relations comprising that structure are directly knowable only from the, subjective, first-person point of view.

It seems arguable then that the key difference between qualia on the one hand and such physical phenomena as functional organization, neurophysiology, and behavior on the other, is that the former are irreducibly subjective, "private," and first-person in character while the latter are inherently objective, publicly accessible, and third-person. The dualist concludes that since the two sorts of phenomena have such irreconcilable essential properties, the former cannot be accounted for in terms of the latter - in which case materialism, ~~which~~ claims that everything real is explicable in terms of objective, third-Person physical phenomena, must be false.

Property dualism

Interestingly, most of the philosophers typically associated with the sorts of arguments surveyed in this chapter are, though critics of main stream materialism, nevertheless not Cartesian dualists. Some of them endorse an agnostic materialism as a fallback position: Joseph levine, for example, suggests that what such arguments really prove is at most that there is an "explanatory gap" between the physical and the mental - that we do not understand how materialism can be true, but that this doesn't show that it isn't true; Colin McGinn adds that it might simply be that evolution has not given us the conceptual resources fully to grasp the manner in which material processes generate mental ones. But such moves arguably miss the point: if the arguments of Chalmers, Jackson, Kripke, et al. work at all, they seem to prove that qualia are just not reducible to physical properties, not that we can't understand how they are reducible. (No one would think it reasonable to reply to Godel's arguments for his famous incompleteness theorems by suggesting that perhaps we just don't understand how the consistency of a formal system containing computable arithmetic is internally provable.)

Most philosophers sympathetic to the arguments in question opt instead for what has come to be known as property dualism, the view (alluded to earlier when discussing the zombie argument) that there is, contrary to Cartesian substance dualism, only one kind of substance - material substance - but that there are also, contrary to materialism, two kinds of properties, physical and non-physical. In this view, the mind, considered as a substance, is indeed identical to the brain, but mental properties - or at least qualia - are not physical properties of the brain, but non-physical properties inhering in its physical substance. The advantage claimed for this view is ihat it can accommodate both the Cartesian dualist's conviction that mind is irreducible to matter and the materialist's insistence that mind is inseparable from matter.

Property dudists also often take other mental phenomena – thosc which don't essentially involve qualia - to be susceptible of explanation in terms of materialistic functionalism in a way qualia are not. This is held to be true in particular of the propositional attiudes - belief, desire, hope, fear - so called because they involve a subject taking a certain attitude toward a proposition, such as the attitude of belief you take toward the proposition that it is raining when you believe that it is raining, or the attitude of hope you take toward the proposition that you will pass your exams when you hope that you will pass your exams. The idea is that since these sorts of mental states are not necessarily associated with qualia (for you could believe that it is raining even if you aren't consciously entertaining the belief at the moment), there is no objection to be made to reducing them to physical states of the brain on the basis of arguments of the "inverted spectrum," "Chinese nation,"zombie' or "knowledge" sort.

Whether this suggestion is as-plausible as Property dualists generally take it to be is something we will explore in chapters 6 and 7. But it might seem to give the property dualist a significant advantage over the Cartesian dualist where defending a broadly dualist view of the world is concerned. As we saw in chapter 2, the Cartesian dualist appears to have a difficult time explaining exactly how a non-physical substance could possibly interact with the body. How, for example, your belief that it is raining can be what causes you to go get your umbrella becomes metaphysically mysterious. Epiphenomenalism looms. But the property dualist might appear to have avoided this problem: your belief is, most

property dualists would allow, a physical state of your brain, so there need be no mystery about how it can have a causal influence on behavior. Even your Perception that it is raining can, in so far as it involves having "propositional attitude as much as a belief does, be identified with a physical process in your brain, so that there is no problem in explaining how it too can cause behavior. True, the perception, unlike many beliefs, may well be associated with certain qualia (such as the sensation of water droplets hitting one's arm), and these cannot be identified with physical properties of the brain. Indeed, it seems that qualia, unlike propositional attitudes, must at the end of the day be regarded as epiphenomenal, playing no role whatever in the production of behavior, since the behavior of a zombie would be exactly the same as that of someone who has qualia. But as long as the perception itself is physical, this shouldn't matter: your perception of the raindrops really does cause you to get your umbrella, even if the qualia associated with it do not.

In fact, however, it matters a great deal, and property dualism seems if anything to have a worse problem with epiphenomenalism than does Cartesian dualism. Recall that the Cartesian dualist who opts for epiphenomenalism seems to be committed to the absurd consequence that we cannot even so much as talk about our mental states, because if epiphenomenalism is true, those mental states have no effect at all on our bodies, including our larynxes, tongues, and lips. But as Daniel Dennett has pointed out, the property dualist seems committed to something even more absurd: the conclusion that we cannot even think about our mental states, or at least about our qualia! For if your beliefs - including your belief that you have qualia - are physical states of your brain, and qualia can have no effect whatsoever on anything physical, then whether you really have qualia has nothing to do with whether you believe you have them. The experience of pain you have in your back has absolutely no connection to your belief that you have an experience of pain in your back; for, being incapable of having any causal influence on the physical world, it cannot be what caused you to have beliefs about it. Indeed, it would also seem to follow that you can have no confidence that the pain even exists in the first place; for you would have exactly the same beliefs about it whether it existed or not. Property dualism thus appears to lead to a skepticism even more radical than that entailed by Descartes's evil spirit scenario: if property dualism is true, then you cannot even be

certain that your-own conscious experiences exist; you might, for all you know, be a zombie!

This is not only bizarre, it is incoherent. The whole point of property dualism is to insist that there are non-physical qualia; if the theory also entails that we can never know that there are such qualia, then how (and why) are we even considering it? How can property dualists themselves so much as formulate their hypothesis? Chalmers attempts to deal with this problem by suggesting that the assumption that there must be a causal connection between the knower and what is known, though appropriate where knowledge of physical objects is concerned, is inappropriate for knowledge of qualia. The existence of a causal chain implies the possibility of error, since (as we saw in chapter l) it seems to entail a gap between the experience of the thing known and the thing itself, a gap between appearance and reality: it ii at least possible that the normal causal chain connecting us to the thing experienced has been disrupted, so that the experience is misleading (as in hallucination or deception by a Cartesian evil spirit). But knowledge of qualia, Chalmers says, is absolutely certain. Here. there is no gap between appearance and reality, because the appearance - the way things seem, which is constituted by qualia themselves is the reality. Knowledge of qualia must therefore somehow be direct and unmediated by causal chains between them and our beliefs about them. The fact that they can have no causal influence on our beliefs thus does not, after all, entail that we can't think or talk about them.

But an objection to this is that it seems question-begging, since whether our knowledge of quaiia really is certain is part of what is at issue in Dennett's argument. Moreover, Chalmers' claim that there is no gap between appearance and reality where knowledge of qualia is concerned seems problematic, given the assumption he shares with other property dualists that propositional attitudes can, unlike qualia, be reduced to physical processes in the brain. For while there is a sense of "appearance" and "seeming" which involves the having of qualia (a sense we can call the "qualitative" Sense), there is also a sense of these words (call it the "cognitive" sense) which does not, but instead involves only the having of certain beliefs: one might say, for example, that at first it seemed or appeared to him that Chalmers' arguments were sound, but on further reflection he concluded that they were not. Here there need be no

qualia present, but only a mistake in judgment or the having of a false belief. But the having of beliefs and the making of judgments are, by Chalmers' own lights, identical with being in certain brain states, so that there is a sense in which even a zombie has beliefs (including false beliefs) and makes judgments (including mistakes in judgment). But in that case, it could "seem" or "appear" even to a zombie that it had qualia, even though by definition it does not. So there can be a gap between appearance and reality even where qualia are concerned. Dennett's challenge remains: how can property dualists so much as think about the qualia they say exist? How can they know that they aren't zombies?

Chalmers' view seems to be that this sort of objection can be avoided by arguing that it is just in the very nature of having an experience that one is justified in believing one has it, that there is a conceptual connection between having it and knowing one is having it. The evidence for my belief that I'm having the experience and the experience itself are the same thing; so I don't infer the existence of the experience from the evidence, but just know directly from the mere having of the evidence. But this seems merely to push the problem back a stage, for now the question is how one can know one really has that evidence - the experience - in the first place, given that an experienceless zombie would also believe that it has it (and, if it's read Chalmers, that there is a conceptual connection between having it and being justified in believing it does). Chalmers' claim seems to amount to the conditional: if you have qualia, then you can know you have them. But that raises the question of how one can know the antecedent of this conditional' i'e' of how one can know one does in fact have qualia. Chalmers' reply is "Because it seems to me that I do, and its seeming that way is all the justification I need." But a zombie would believe the same thing! "But I have evidence the zombie doesn't have - my experience!" Chalmers would retort. Yet the zombie believes that too, because it also seems to it (in the cognitive sense) that it has such evidence. Any response Chalmers could, give to such questions would seem to invite further questions about whether he really has the evidence he thinks he does. His only possible reply can be to say that he has it because he seems to have it, but if he says that he seems to in the cognitive sense of "seems," then he's saying something even a zombie would believe, while if he says, even to himself, that he seems to in the qualitative sense of "seems," then he's

begging the question, for whether he has the qualia that this sense of "seems" presupposes is precisely what's at issue. Chalmers' reply to the sort of criticism raised by Dennett thus seems to fail.

Property dualism would thus appear to lead to absurdity as long as it concedes to materialism the reducibility of the propositional attitudes. If it instead takes the attitudes to be, like qualia, irreducible to physical states of the brain, this absurdity can be avoided: for in that case, your beliefs and judgments are as non-physical as your qualia are, and there is thus no barrier (at least of the usual mental-to-physical epiphenomenalist sort) to your qualia being the causes of your beliefs about them. But should it take this route, there seems much less motivation for adopting property dualism rather than full-blown Cartesian substance dualism: it was precisely the concession of the materiality of propositional attitudes that seemed to allow the property dualist to make headway on the interaction problem, an advantage that is lost if that concession is revoked; and while taking at least beliefs, desires, and the like to be purely material undermines the plausibility of the existence of a distinct non-physical mental substance, such plausibility would seem to be restored if all mental properties, beliefs and desires, as much as qualia, are non-physical. Moreover, property dualism raises a puzzle of its own, namely that of explaining exactly how non-physical properties could inhere in a physical substance.

Property dualism, then, is arguably not a genuine advance over substance dualism, though some of the arguments of property dualists appear to pose a significant challenge to materialism and thereby to advance the cause of dualism generally. Yet the materialist still has the interaction problem to wield against the dualist, along with the less paradoxical but still unsatisfactory form of epiphenomenalism that threatens even Cartesian dualism. Moreover, the materialist's last word about qualia has not yet been spoken. We've seen that the problem qualia pose for the materialist is, at bottom, the problem of accounting for the existence of a conscious subject having a first-person point of view on the world. An adequate understanding of the qualia problem cannot be had, ~~then~~, unless it is considered as part of the broader problem of the nature of consciousness itself. If consciousness in general can be explained in entirely materialistic terms, maybe a materialist account of qualia in particular would be possible after all, as a by-product of this

more general theory. That, at any rate, is the hope of a number of contemporary materialist philosophers. A look at the problem of consciousness must therefore be the next item on our agenda.

Further reading

Block's Chinese nation scenario is from his "Troubles with Functionalism," reprinted in both the Rosenthal and Chalmers anthologies cited at the end of the last chapter. Jackson's version of the knowledge argument is presented in "What Mary Didn't Know" and Nagel's in "What is it Like to Be a Bat?"; Churchland's reply is in "Knowing Qualia: A Reply to Jackson," Lewis's in "What Experience Teaches," and van Gulick's in "Understanding the Phenomenal Mind: Are We All Just Armadillos?" These essys have been reprinted in numerous places (some of them in the Chalmers and Rosenthal anthologies), but they can all be found together in Ned Block, Owen Flanagan, and Guven Guzeldere, eds. The Nature of Consciousness: Philosophical Debates (Cambridge, MA: MIT Press,1997). Sydney Shoemaker's "The Inverted Spectrum," which discusses that famous thought experiment, can also be found in this anthology. The structure of color space is the subject of C. L. Hardin's Color for Philosophers (Indianapolis: Hackett, 1988) andAusten Clark's Sensory Qualities (Oxford: Clarendon Press, 1993). Carnap's analysis is in his classic Der logische Aufbau der Welt, translated by R. George as The logical Structure of the World (Berkeley and Los Angeles: University of California Press, 1967).The section of Kripke's Naming and Necessity defending the modal or zombie argument (though not by that name) is reprinted in the Chalmers, Rosenthal, and Block et al. anthologies; Levine's "Materialism and Qualia: The Explanatory Gap" and McGinn's "Can We Solve the Mind-Body Problem?" are also available in the Chalmers anthology. Chalmers defends property dualism and the zombie argument at great length in The Conscious Mind (New York: Oford University Press, 1996). Dennett's critique of properry dualism is in his consciousness Explained (Boston, MA: Little, Brown, and Company, I99l). That subjectivity rather than intrinsicality is the core of the concept of qualia is a thesis I defended earlier in "Qualia: Irreducibly

Subjective but not lntrinsic," journal of Consciousness Studies, Vol. 8, No. 8 (August 2001).