In any case, our best remaining option is reductionism. Does this mean that we are committed willy-nilly to reductionism? The answer is no: what we have established, if our considerations have been generally correct, is a conditional thesis, "If mentality is to have a causal influence in the physical domain—in fact, if it is to have any causal efficacy at all—it must be physically reducible." I have not argued for reductionism simpliciter; rather, I have argued that mental causation requires reduction, and that anyone who believes in mental causation must be prepared to endorse mind-body reduction. We may call this "conditional reductionism." It is important to keep in mind that this is not reductionism tenet curt. Moreover, none of this says anything about the truth or plausibility of reductionism. Whether or not the mental can be reduced to a physical base is an independent question that must be settled on its own merits. Those of us who believe in mental causation should hope for a successful reduction. But again this is only a wish; it doesn't make reducibility real or reductionism true.8

So we have finally come to a crossroads: Can we physically reduce minds? Is mentality reducible in physical terms?

Reducing Minds

In raising this question about the reducibility of the mental, it is important not to think that the mental as a totality must be either all reducible or all irreducible. It may well be that parts of the mental are reducible while the rest is not. It may be that

8. I have sometimes been described as a reductionist (for example, Robert Van Gulick refers to me as a "hard core reductive materialist" in his survey article "Reduction, Emergence and Other Recent Options on the Mind-Body Problem: A Philosophical Overview," Journal of Consciousness Studies 8 (2001): 1-14, p. 2. I am sure I have often written and spoken in confusing ways in the past, but I hope this sets the record straight.
the physicalist project can be carried through for various subdomains of the mental but not for all, and that physicalism can be vindicated for much of the world but not for the whole world. I will argue that this is in fact the case, namely that much of the mental domain can be physicalized but not all of it. More specifically, my view is that the qualitative characters of conscious experience, what are now commonly called “qualia,” are irreducible, but that we have reason to think that the rest, or much of it anyway, is reducible. I am of course not the first, or only, person to hold a view like this. Joe Levine, David Chalmers, and others have argued, quite plausibly in my opinion, for just such a position.

Chalmers distinguishes between two classes of mental states, those he calls “psychological” and those he calls “phenomenal.” Psychological states are states that “play the right sort of causal role in the production of behavior,” and include states like belief, desire, memory, and perception. I will call them “intentional/cognitive properties.” Phenomenal states are states with a qualitative character—or, to use a phrase popularized by Thomas Nagel, states such that there is “something it is like” to be in those states—like pain, itch, the visual experience of yellow when you look at a field of sunflowers, and the tactile sensation you experience when you run your fingers over a smooth marble surface. On Chalmers’s view, intentional/cognitive properties are physically reducible, whereas phenomenal properties, or “qualia,” resist reduction.

But what is it to “reduce” something? What needs to be done to accomplish a physical reduction of something mental? Consider the gene and how it has been reduced in molecular biology. The concept of a gene is the concept of a mechanism in an organism that encodes and transmits genetic information. That, I believe, was indeed the concept that Mendel, the founder of modern genetics, had in mind when he spoke of “genetic factors.” Genetic factors were to be whatever mechanisms or processes in organisms were causally responsible for the transmission of heritable characteristics. In short, the concept of a gene is defined in terms of a causal function, or causal role—in terms, that is, of the causal task that must be performed by whatever it is that is to qualify as a gene. As we will say, the concept of a gene is a “functional” concept, and the property of being a gene is a functional property defined by a “job description.” A functional conception of the gene defines the scientific research program: Identify the mechanisms—say, in pea plants or fruit flies or whatever—that perform the task of transmitting heritable characteristics. Research in molecular genetics has shown, we are told, that it is DNA molecules that perform this task—they are the genes we have been looking for. That is to say, the DNA molecules are the “realizers” of the gene. Moreover, molecular genetics provides us with an explanation of how DNA molecules manage to perform this complex causal work. When all this is in, we can say that the gene has been physically reduced, and that we now have a reductive explanation of how the process of heredity works at the molecular level. Notice that while, as far as we know, DNA molecules are the genes for terrestrial life, there is no reason to presume that they are the only possible physical mechanisms capable of performing the causal tasks associated with the concept of a gene. It might well be that in certain extraterrestrial organisms these tasks are performed.

11. Chalmers would put this in terms of reductive explanation, not reduction. However, his concept of reductive explanation closely corresponds to my notion of reduction; on the functional model of reduction that I favor, reductive explanation is not separable from reduction. See chapter 4 for details.
12. Reduction was discussed earlier in greater detail; see chapters 1 and 4.
not by DNA molecules but by molecules of another kind, say XYZ; for them, XYZ molecules, not DNA molecules, would be the genes. This sort of “multiple realization” must be expected, but the phenomenon of multiple realization, whether actual or only nomologically possible, does not, as some philosophers used to assume, represent an impediment to reduction or reductive explanation.13

To summarize, then, reduction can be understood as consisting of three steps: The first is a conceptual step of interpreting, or reinterpreting, the property to be reduced as a functional property, that is, in terms of the causal work it is supposed to perform. Once this has been done, scientific work can begin in search of the “realizers” of the functional property—that is, the mechanisms or properties that actually perform the specified causal work—in the population of interest to us. The third step consists in developing an explanation at the lower, reductive level of how these mechanisms perform the assigned causal work. When the first step has been carried out and the property targeted for reduction has been functionalized, in an important sense the property has been shown to be “reducible”—it is now a matter of scientific research to find the realizers. That is, if anything has the functionalized property, it follows that it instantiates some lower-level physical realizer, and it must in principle be possible for scientific investigation to identify it. Even if we have not identified the actual realizer—perhaps we never will—it would make not much difference philosophically: we know that there must be a lower-level physical realizer, even if we don’t have a perspicuous description of it in an underlying theory, and we know the phenomenon involved to be reducible to its physical realizer, whatever it is. This means that as far as the metaphysical situation is concerned the functional definability of a property is the only issue that matters. That a property is functionalizable—that is, it can be defined in terms of causal role—is necessary and sufficient for functional reducibility. It is only when we want to claim that the property has been reduced (for a given system) that we need to have identified its physical realizer (for that system).

Our question about reduction of minds, therefore, comes to this: Are mental properties functionalizeable? Can they be defined or characterized in terms of their causal work? The answer, as I have indicated, is yes and no. No for qualitative characters of experience, or “qualia,” and yes, or probably yes, for the rest.

Why should we think that intentional/cognitive properties like believing, desiring, and intending are functionally definable, in terms of the work they do? I do not believe that anyone has produced full functional definitions for them, and it is perhaps unlikely that we will have such definitions any time soon. However, there are various considerations that indicate that these properties are functional properties and should be characterizable in terms of the causal work they do in the overall economy of human behavior. Consider a population of creatures, or systems, that are functionally and behaviorally indistinguishable from us, and, in general, observationally indistinguishable from us. Exact indistinguishability or indiscernibility is unimportant; we may simply suppose that their behaviors are largely similar to ours. They interact with their physical environment, which we may suppose is pretty much like ours, and interact with one another, much the same way we do. In particular, they exhibit similar linguistic behavior; as far as we can tell, they use language as we do for expressive and communicative purposes. If all this is the case, it would be incoherent to withhold states like belief, desire, knowledge, action, and intention from these creatures. If, for example, we grant that these creatures are language users, that alone would

be sufficient to qualify them as creatures with thought, belief, understanding, intentionality, meaning, and the rest. Assertion is fundamental to speech, and a language user must be capable of making assertions. When someone makes an assertion, he expresses a belief. When someone asks a question or gives a command, he is expressing a thought and a desire. A language user is a cognitive being with full intentionality. And if we grant agency to these creatures (how could we not, given that they have belief and desire, and that their observable behavior is like ours?), we will be compelled to see them as creatures with desires, preferences, and intentions, and whose behavior and actions can be evaluated according to the norms of rationality. It seems to me that we cannot avoid thinking of intentional/cognitive states, like thought, belief, and desire, as supervenient on behavior and other observable physical facts. We must accept creatures that are behaviorally and functionally like us as creatures with a mentality similar to ours—with belief, desire, intentionality, will, and so on. This is one strong reason for thinking that such mental properties are definable and interpretable in terms of their roles in behavior causation.

Looking at the situation less globally, suppose that we are told to create a device that perceives shapes and colors of medium-sized objects presented to it (perception), processes and stores the information so gained (belief, memory, knowledge), and uses it to guide its actions (agency). I believe we know how to go about designing and building machines with such capacities; in fact, I believe simple machines with such powers have already been manufactured. That is because these states and processes, like perception, belief, memory, and using information to guide action, are specifiable in terms of their causal roles, or “job descriptions.” A creature, or system, that has the capacity to do certain things in certain ways under certain conditions is ipso facto something that perceives, remembers, and appropriately behaves. I believe this is why mental talk comes so naturally in describing the activities and capacities of certain artifacts; we talk of a radar system “mistaking” a weather balloon for an approaching airliner, a chess-playing computer as “trying to capture” an enemy pawn, and even the humble supermarket automatic door opener as “seeing” or “knowing” that a customer is approaching. Mental talk is even more natural and familiar for animals—not only dogs and cats, but also those farther removed from us, say, mollusks and insects and even amoebas.

These are among the reasons for thinking that cognitive/intentional mental properties are closely tied, conceptually and semantically, to behavior. This does not mean that we are now, or ever will be, in a position to produce neat functional definitions for complex and multifaceted capacities and properties like belief, desire, and emotion. Logical behaviorists and functionalists are widely thought to have failed to deliver such definitions. But that is not crucial. I think the following two facts are important in this context: First, partial functional analyses of these properties in terms of their causal work, even if complete analyses are not available, can get us going with the scientific projects of searching for the underlying physical/biological mechanisms. We don’t have to know all the things that belief does before we start work on uncovering its possible neural mechanisms; a partial list will be enough to start us off. The list will grow richer and more detailed, and this will provide directions for further scientific work. Second, the fact is that even though a complete analysis of belief is not in and perhaps never will be, we don’t think there is ultimately anything beyond causal work vis-à-vis observable behavior that is involved in belief. To be sure, beliefs can generate further beliefs; in conjunction with desires, they can cause further desires; and so on. However, these further mental states, too, must ultimately be anchored, conceptually and epistemologically, in observable behavior. In looking for causal mechanisms that ground beliefs, these causal connections with other mental states must be accounted for and identified. This only means
that the identification of the realizers of belief must go hand in hand with the identification of realizers of these other mental states. And research results at the level of realizers may lead to the reshaping of the higher-level mental concepts in various ways. As far as intentional states are concerned, we are within the domain of behavior and the physical mechanisms involved in their production; they do not take us outside this domain. 14

Let us now turn to sensory states. Can we reduce qualitative states of consciousness? 15 Suppose we are given another engineering project. This time, we are asked to design a machine that responds to punctures and abrasions to its own skin ("tissue damage") by taking evasive maneuvers to separate itself from the source of the damage ("escape behavior"); in addition, we are told to make this device experience pain when it suffers damage to its skin. That is, we are asked to design into the machine a "pain box" which, in addition to its causal work of triggering an appropriate motor response when it suffers damage, gives rise to a pain experience. We can, I am sure, easily design into a machine a device that will serve as a causal intermediary between the physical input and the behavior output, but making it experience pain is a totally different affair. I don't think we even know where to begin. What we miss, something that we need to know in order to design a pain-experiencing machine, is a connection between the causal work of the pain box and the arising of pain when the box is activated. Why pain rather than itch or tickle? The machine would try to flee when its skin is punctured even if we had, unwittingly or unwittingly, designed itch or tickle into the box.

What this shows is that we cannot distinguish pain from itch or other sensations by their causal work; our strong intuition is that even if pain is associated with scratching behavior (like itch) or squirting behavior (like tickle), as long as it is felt as pain—as long as it hurts—it is pain. Pain may be associated with certain causal tasks, but these tasks do not define or constitute pain. Pain as a sensory quale is not a functional property. In general, qualia are not functional properties. As far as we now know, the only way to create a system with conscious experience is to duplicate an appropriate animal or human brain.

Some philosophers have invoked the conceivability of "zombies" to show that qualia are not logically supervenient on physical/biological facts. Zombies are creatures that, though physically (hence also functionally and behaviorally) indistinguishable from us, have no conscious experience. The zombie hypothesis has been controversial; 16 less controversial is the quale inversion hypothesis, namely the possibility of creatures like us, perhaps other human beings, whose quality space is inverted with respect to ours—who, for example, when they look at mounds of lettuce, experience a color quale of the kind we experience when we look at ripe tomatoes, and who, when they look at ripe tomatoes, sense the color that we sense when we look at lettuce. Such spectrum-inverted people would be as

14. This can be considered a reply to Block and Stalnaker, who argue, in "Conceptual Analysis, Dualism, and the Explanatory Gap," that functional reduction is hopeless because functional definitions of psychological states are not available and never will be. See chapter 4.

15. I have already argued for a negative answer (in chapter 1); here we will briefly go over the same ground but introduce a few new considerations, though no new argument.

16. In fact, I believe the zombie hypothesis is untenable; Chalmers-style zombies are not conceptually possible. My reason is based on what Chalmers calls "the paradox of phenomenal judgments." Zombies are indistinguishable from us in their speech behavior, and we must regard them as genuine language users. Among the assertions they make are "My elbow hurts." "This mosquito bite is really itchy." and the like; they make phenomenal assertions of the sort we make, and do so under similar conditions. Moreover, their phenomenal assertions are not easily isolated; they are integrated smoothly and seamlessly with other parts of their discourse. To hold onto the zombie hypothesis, we must apply a massive "error theory" to these creatures—namely that all their (positive) phenomenal assertions are false. I believe this is incoherent. We must grant that the creatures lack inner consciousness, although the qualitative character of their consciousness remains undetermined.
adaptable as we are in picking tomatoes out of mounds of lettuce and obeying traffic signals, and in general they would do just as well as we do with any other tasks requiring discrimination of red from green. If this is the case, color qualia do not supervene on behavior; two perceivers who behave identically with respect to input applied to their sensory receptors can have different sensory experiences. If that is true, qualia are not functionally definable, they are not task-oriented properties.

So qualia are not functionalizable, and hence physically irreducible. Qualia, therefore, are the “mental residue” that cannot be accommodated within the physical domain. This means that global physicalism is untenable. It is not the case that all phenomena of the world are physical phenomena; nor is it the case that physical facts imply all the facts. There is a possible world that is like this world in all respects except for the fact that in that world qualia are distributed differently. I don’t think we can show it to be otherwise.

**Living with the Mental Residue**

So what do we do with this mental residue? If we want to keep mental causation, we should try to minimize its scope and impact as far as possible. Can the antiphysicalist celebrate his victory? Hardly. For one thing, the mental residue encompasses only qualitative states of consciousness, and does not touch the intentional/cognitive domain. And it is in this domain that our cognition and agency are situated. Second, ordinary sensory concepts, like “pain,” “itch,” and the rest, have motivational/behavioral aspects in addition to qualitative/sensory aspects, and it is clear that the motivational/behavioral component of, say, pain can be given a functional account. Third, we will do well to remember that our conditional reductionism still stands: If anything is to exercise causal powers in the physical domain, it must be an element in the physical domain or be reducible to it. This has two direct implications.

First, the mental residue, insofar as it resists physical reduction, remains epiphenomenal. It has no place in the causal structure of the world and no role in its evolution and development. Second, if we are right about the reducibility of cognitive/intentional mental states, we have vindicated their causal efficacy—and thereby largely, if not completely, solved the problem of mental causation. Consider Fodor’s lament over the possible loss of mental causal efficacy:

> if it isn’t literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for saying . . . , if none of that is literally true, then practically everything I believe about anything is false and it’s the end of the world.

Three mental causes are on Fodor’s wish list: wanting, itching, and believing. We have good news for Fodor—his world is not coming to an end, at least not completely, because two items on his list, wanting and believing, turn out to be in good shape. Two out of three isn’t bad! But what can we tell Fodor about itching? Should he care about itching, as much as about wanting and believing? At least we can say this: If we can save intentional/cognitive properties, we can save our status as cognizers and agents. Saving itching isn’t required for saving cognition or agency.

Actually, though, I believe we can go some distance toward saving qualia, though not all the way. I have earlier noted how two persons whose color spectra are inverted with respect to each other can exhibit the same discriminative behavior. Brief reflection shows that some important aspects of qualia are

---

quite directly manifestable in behavior and therefore functionalizable. For analogy, consider traffic lights: everywhere in the world, red means stop, green means go, and yellow means slow down. But that is only a convention, the result of a social arrangement; we could have adopted a system according to which red means go, green means slow down, and yellow means stop, or any of the remaining combinations. That would have made no difference to traffic management. What matters are the differences and similarities among colors, not their intrinsic qualities. In fact, we could have chosen shapes instead of colors, with circle meaning go, square meaning stop, and so on. Discrimination is what matters; qualities discriminated do not. I believe Moritz Schlick made the observation that what can be communicated about experiences is their form, not their content.

Suppose that we have already acknowledged that a given perceiver can experience a range of qualia. When we present to her a ripe tomato, we may not know, and may not care, what the intrinsic quality of her visual experience is—what color quale she is experiencing. Similarly, when we present to her a bunch of spinach leaves, we may not know what quale characterizes her visual experience. However, we can tell whether her color quale of the tomato is the same as, or different from, her color quale of the spinach leaves. When we next present to her a head of lettuce, we can tell whether the quale she is now experiencing is similar to, or different from, each of the two color qualia she has just experienced. That is, the intrinsic qualities associated with qualia are, or may be, undetectable, but differences and similarities between qualia, within a single individual, are behaviorally detectable, and this opens a way for their behavioral functionalization.\footnote{18. On these issues, see Sydney Shoemaker, “The Inverted Spectrum,” \textit{Journal of Philosophy} 74 (1977): 357–81. What I am saying here, I believe, is in line with Shoemaker’s view that although “absent qualia” are not possible, qualia inversion is metaphysically possible.}

And intuitively that seems right. The fact that blue looks just \textit{this} way to me, green looks \textit{that} way, and so on, should make no difference to the primary cognitive function of my visual system—its function in the generation of information about the physical environment of the sort that makes a difference to my survival and flourishing. Color-inverted persons, as long as they have the capacity to make the same color discriminations, should do as well as we do in learning about the world and coping with it. Intrinsic qualities of qualia are not functionalizable and therefore are irreducible, and hence causally impotent. They stay outside the physical domain, but they make no causal difference and we won’t miss them. In contrast, certain important relational facts about qualia, in particular their similarities and differences, are detectable and functionalizable, and can enjoy causal powers as full members of the physical world. But there is a further question: Why are there such things as qualia? Because we need them \textit{as} place markers, \textit{without} them there \textit{cannot be} no quale differences or similarities. Without content, there can be no form, no structure. You may now ask: Why are there just these qualia and not other possible ones? That remains a mystery; I do not believe that the present approach is capable of answering that question.

\textbf{Where We Are at Last with the Mind-Body Problem}

I feel that the position I have been describing here is a plausible terminus for the mind-body debate. There are many issues that need to be sorted out in more detail and with greater care and precision; among them are the functional reducibility of cognitive/intentional states, the functionalizability of qualia differences and similarities, whether qualia epiphenomenalism is consistent with the assumed fact that the subject of experiences is cognitively aware of them and is able to make
reports about them, and the question whether it is possible to combine qualia epiphenomenalism with full causal efficacy of qualia similarities and differences. But in spite of the further work required, I feel that the remaining work is for the most part a mopping-up operation, and that the important outlines of the position stand out with clarity and salience.

So here is the position that has emerged. It begins by embracing ontological physicalism. Taking mental causation seriously, it also embraces conditional reductionism, the thesis that only physically reducible mental properties can be causally efficacious. Are mental properties physically reducible? Yes and no: intentional/cognitive properties are reducible, but qualitative properties of consciousness, or "qualia," are not. In saving the causal efficacy of the former, we are saving cognition and agency. Moreover, we are not losing sensory experiences altogether: qualia similarities and differences can be saved. What we cannot save are their intrinsic qualities—the fact that yellow looks like this, that ammonia smells like that, and so on. But, I say, this isn't losing much, and when we think about it, we should have expected it all along.

The position is, as we might say, a slightly defective physicalism—physicalism marqué but not by much. I believe that this is as much physicalism as we can have, and that there is no credible alternative to physicalism as a general worldview. Physicalism is not the whole truth, but it is the truth near enough, and near enough should be good enough.

References


