8.3 Property Dualism and Subjective Experience

The nature of subjective experience has seemed to many people so striking and so extraordinary that it has been invoked repeatedly as the standing refutation of reductionism. The argument from subjective experience has been most powerful, not in the hands of the substance dualists, who have to contend with complications of their ghostly substance, but in the hands of the property dualists. Although there are nontrivial differences among the hypotheses advanced by assorted property dualists, the crux of the shared conviction is that even if the mind is the brain, the qualities of subjective experience are nevertheless emergent with respect to the brain and its properties. Subjective experience, goes the argument, has a character and a quality uniquely and irreducibly mental.

Since the notion of a property’s being emergent makes an appearance in this argument, an explication of “emergence” is in order. In general, whether a property is emergent is a function of the reductive relation that holds—or rather, fails to hold—between two theories or conceptual frameworks. More specifically, a property $P$ specified by its embedding theory $T_1$ is emergent with respect to the properties of an ostensibly reducing theory $T_2$ just in case

1. $P$ has real instances,
2. $P$ is co-occurrent with some property or complex feature recognized in $T_2$, but nevertheless
3. $P$ cannot be reduced to any property postulated by or definable within $T$ (Paul M. Churchland 1985)

As noted in the account of intertheoretic reduction outlined in chapter 7, the reducibility of one property to another depends on whether the theory that characterizes the property at issue reduces to the theory that characterizes the other. To put the matter informally, if a property of one theory has causal powers that are not equaled or comprehended by any property in the second, more basic theory, then the property is considered to be emergent with respect to the second theory. Accordingly, to claim that the qualitative features displayed in one’s subjective experience are emergent with respect to the physical brain is to insist that the commonsense conceptual framework for apprehending and describing such psychological properties is not reducible to any future neuroscience.

Whether a property is emergent is therefore not a simple observational feature of the property, and so one cannot tell simply by inspecting a property whether or not it is emergent with respect to some other property, despite the conviction displayed by the occasional theorist lost in introspective reverie. (E. Nagel (1961) also points this out.) Nor of course do commonsense intuitions that two properties are substantially or even stunningly different entail anything about whether a future intertheoretic reduction might actually identify the two. Light may seem completely different from electromagnetic radiation, yet light turns out to be electromagnetic radiation. Having a high temperature seems supremely different from having a high mean molecular kinetic energy, yet it turns out that high temperature in a gas is high mean kinetic energy of the constituent molecules. Notice also that one does not provide independent evidence for the irreducibility of one property to another merely by claiming that the first is emergent relative to the second. That would be like saying of a property that it is irreducible because it is irreducible.

“Emergent property” is also used in the neuroscientific literature with a quite different sense roughly equivalent to “network property.” Consider a set of cells in the retina that are wired so as to collectively constitute a movement detector, even though none of the individual cells is itself a movement detector. The functional property of being a movement detector may understandably be described as “emergent” relative to the individual neurons in the circuit. How-
ever, the functional property is certainly and obviously reducible to the neurophysiological properties of the network. Indeed, once we understand the network, we have the reductive story in hand. Although this is a useful sense of “emergence” (which Dennett calls "innocent emergence"), it is clearly not the sense intended by property dualists in their arguments against reductionism. Thus, when Sperry (1980) argues that mental states are emergent, he specifies that he means they are irreducible, not merely that they are network properties.

The claim that subjective experience is not reducible to brain states is to be understood within the wider framework of intertheoretic reduction, where it unpacks as the claim that psychology will not reduce to neuroscience; more specifically, that it will not reduce to neuroscience in such a way that subjective experiences can be identified with states of the brain. States of the brain are causally connected to subjective experience and give rise to the stream of events in awareness, the argument will agree, but the experience itself, with its unique qualities, cannot be identified with some process or aspect of neuronal activity. In contrast to the substance dualists, the property dualists do not believe there is a nonphysical substance in which experiences inhere. Rather, they claim that subjective experiences are produced by the brain and can in their turn affect the brain, but they are not themselves identifiable with any physical properties of the brain. On this view, we cannot say, for example, that feeling sad is a neuronal configuration in such and such a neuronal ensemble.

An analogy here may help. It has often been claimed that the blueness of (liquid) water is a property that is emergent relative to the microphysics of H$_2$O molecules, on the grounds that no amount of microphysical information could allow us to predict or to deduce that liquid aggregates of such molecules would have the peculiar qualitative character we call “blue.” Blueness may systematically co-occur with aggregates of H$_2$O molecules, it is conceded, but it is a self-contained and irreducible property that appears in addition to the microphysical features of aggregated H$_2$O. It is, in a word, emergent.

Given the account of intertheoretic reduction outlined earlier, the fallacy of this reasoning is displayed with relative ease. For one thing, reduction does not require that reduced properties, as conceived within their older conceptual framework, be deductible or predictable from within the new reducing theory. Old frameworks are culturally idiosyncratic and highly various. It cannot be the obligation of new theories to predict how ignorant cultural epochs may happen to conceive of the complex phenomena in their explanatory domain. What they are obliged to do, if they are to achieve the reduction of earlier concepts, is nothing more and nothing less than to entail the existence of properties that systematically mimic the alleged causal powers of the properties to be reduced.

In the example at issue, the microphysics of H$_2$O molecules does indeed entail that liquid aggregates of them will preferentially scatter incident electromagnetic radiation at a wavelength of about 0.46 μm. It is this complex property that proves to have all of the causal powers of blueness (at least as it is manifested in liquid water). This microphysical property affects human observers in all the same ways as does blueness. It affects nonhuman instruments in all the same ways as does blueness (for example, it projects light through a prism into the same spectral position as does blueness). And so forth. Because of these systematic parallels, it is reasonable to identify the blueness of an object with its disposition to scatter (or reflect, or emit) electromagnetic waves preferentially at about 0.46 μm. That is the property that humans have been visually discriminating for millennia, though without appreciating its fine-grained nature. The blueness of water, therefore, is not an emergent property. On the contrary, it reduces rather smoothly, and as a coherent part of a systematic account in which the other colors are also reduced. (For simplicity's sake, I here ignore more recent accounts according to which color is a matter of the object's reflectances at, not one, but three critical wavelengths. Those new accounts are more complex, but the reductive lesson is the same.)

We must not be stampeded, therefore, either into accepting impossibly strong conditions on successful reduction or into forgetting the point of intertheoretic reduction, which is just to show that a new theoretical framework provides a parallel but markedly superior conception of an old and familiar domain. Neuroscience should not be required to meet standards of reduction more stringent than we impose elsewhere in science simply because the concepts it aspires to reduce are deeply entrenched in common sense.

Furthermore, it is true that at this stage in the history of science, subjective experience has not been successfully identified with and explained by states and processes in the brain. Insofar as it has not, then subjective experiences do not now enjoy a reduction to brain states. Nevertheless, this current state of science does not entail that no neuroscientific theory will ever reduce psychology. Analogously, in the eighteenth century it would have been true to say that temperature did not reduce, inasmuch as thermodynamics was still an autonomous science. This did not entail that it never would reduce to a more basic theory or that temperature is an emergent property. In-
deed, by the late nineteenth century thermodynamics was the beneficiary of a triumphant reduction to statistical mechanics, at which time it was evident that temperature (of gases) is not emergent but is identical with mean molecular kinetic energy.

Now in saying that subjective experience is emergent, property dualists wish to claim not merely that the reduction of folk psychology to neuroscience eludes us here and now, but that it always will. In principle. Evidently this is much stronger than simply saying that psychology has not yet been reduced to neuroscience, which, after all, is a crashingly obvious thing to say. How does the property dualist propose to defend the opinion that neuroscience will never reduce psychology in such a way that subjective experience can be identified with states of the brain?

The main arguments derive from reflections on the special nature of subjective experiences. But the objections are not always either clear or well-defined, and in many cases one cannot tell what it is about subjective experiences that is to justify the prediction that neuroscience can never reduce psychology. Sometimes the arguments simply work fuzzy intuitions about what is and is not imaginable, or even about what is and is not desirable. Since it is impossible here to catalogue all the variations on this theme, let alone analyze them, instead I have selected for further discussion what I take to be the strongest and most coherent of the arguments. The arguments are extracted from Thomas Nagel’s classic paper (1974) and Frank Jackson’s more recent but destined-to-be classic paper (1982). For the analysis and the criticism of these arguments, I have made extensive use of two works by Paul M. Churchland (1984, 1985).

(A)

(1) The qualia of my sensations are knowable to me by introspection.
(2) The properties of my brain states are not knowable to me by introspection.
Therefore:
(3) The qualia of my sensations ≠ the properties of my brain states.

A second argument, complementary to the first, seems also in play:

(B)

(1) The properties of my brain states are knowable by the various external senses.
(2) The qualia of my sensations are not knowable by the various external senses.
Therefore:
(3) The qualia of my sensations ≠ the properties of my brain states.

The general form of the argument seems to be this:

(1) a is F
(2) b is not F
Therefore:
(3) a ≠ b

Leibniz’s law says that a = b if and only if a and b have every property in common. So if a = b, then if a is red, b is red, if a weighs ten pounds, then b weighs ten pounds, and so forth. If a is red and b is not, then a ≠ b. Assuming their premises are true, arguments (A) and (B) appear to establish the nonidentity of brain states and mental states. But are their premises true?

Let us begin with argument (A). There is no quarrel with the first premise (the qualia of my sensations are known-to-me-by-introspection), especially since qualia are defined as those sensory qualities known by introspection, and in any case I have no wish to deny introspective awareness of sensations. In contrast, the second premise (the properties of my brain states are not known-to-me-by-introspection) looks decidedly troublesome. Its first problem is that it begs the very question at issue—that is, the question of whether or not mental states are identical to brain states. This is easy to see when we ask what the justification is for thinking that premise true.

The point is this: if in fact mental states are identical to brain states, then when I introspect a mental state, I do introspect the brain state...
with which it is identical. Needless to say, I may not describe my mental state as a brain state, but whether I do depends on what information I have about the brain, not upon whether the mental state really is identical to some brain state. The identity can be a fact about the world independently of my knowledge that it is a fact about the world. Similarly, when Jones swallows an aspirin, he thereby swallows acetylsalicylic acid, whether or not he thinks of himself thus; when Oedipus kissed Jocasta, he kissed his mother, whether or not he thought of himself thus. In short, identities may obtain even when we have not discovered that they do. The problem with the second premise is that the only justification for denying that introspective awareness of sensations could be introspective awareness of brain states derives from the assumption that mental states are not identical with brain states. And that is precisely what the argument is supposed to prove. Hence the charge of begging the question. (Although I have used (A) as an illustration, the same kind of criticism applies equally to (B).)

Other problems with these arguments are more subtle. One difficulty is best brought out by constructing an argument analogous to (A) or (B) with respect to the character of the properties under discussion and comparing the arguments for adequacy. Consider the following arguments:

(C)
(1) Smith believes Hitler to be a mass murderer.
(2) Smith does not believe Adolf Schicklgruber to be a mass murderer.
Therefore:
(3) Adolf Schicklgruber ≠ Adolf Hitler.

As it happens, however, Adolf Schicklgruber = Adolf Hitler, so the argument cannot be right.

Or consider another instance of the general argument form where the property taking the place of F is a complex property concerning what John believes or knows:

(D)
(1) Aspirin is known by John to be a pain reliever.
(2) Acetylsalicylic acid is not known by John to be a pain reliever.
Therefore:
(3) Aspirin ≠ acetylsalicylic acid.

And one final example more closely analogous to the arguments at issue:

(E)
(1) Temperature is directly apprehendable by me as a feature of material objects.
(2) Mean molecular kinetic energy is not directly apprehendable by me as a feature of material objects.
Therefore:
(3) Temperature ≠ mean molecular kinetic energy.

These arguments fall because being-recognized-as-a-something or being-believed-to-be-a-something is not a genuine feature of the object itself, but rather is a feature of the object as apprehended under some description or other or as thought about in some manner. Having a certain mass is a property of the object, but being-thought-by-Smith-to-have-a-certain-mass is not a genuine property of the object. Such queer properties are sometimes called "intentional properties" to reflect their thought-mediated dependency. Notice that in (B) the property is being-known-by-the-various-external-senses, and in (A) the property is being-known-by-me-by-introspection. Both are sterling examples of thought-dependent properties.

Now the arguments (C) through (E) are fallacious because they treat intentional properties as though they were genuine properties of the objects, and a mistake of this type is called the intentional fallacy. It is evident that the arguments designed to demonstrate the nonidentity of qualia and brain states are analogous to arguments (C) through (E). Consequently, they are equally fallacious, and the nonidentity of mental states and brain states cannot be considered established by arguments such as (A) and (B).

The last difficulty with the arguments is better seen in a slightly different and more compelling version of the argument for the nonidentity of mental states and brain states, which I present and discuss below.

Knowing Our Sensations: Jackson's Argument
The strategy of this second argument once again involves showing that differences between knowing our states via introspection and knowing via nonintrospective means are of such a nature as to constitute grounds for denying the reducibility of psychology to neuroscience. In order to clarify those differences, Frank Jackson (1982) has constructed the following thought-experiment. Suppose that Mary is a neuroscientist who has lived her entire life in a room carefully controlled to display no colors, but only shades of white, gray, and black. Her information about the outside world is transmitted to her by means of a black-and-white television. Suppose further that one
way or another she comes to know everything there is to know about the brain and how it works. That is, she comes to understand a completed neuroscience that, among other things, explains the nature of thinking, feeling, and perception, including the perception of colors. (This is all wildly unlikely, of course, but just suppose.)

Now for the argument: despite her knowing everything there is to know about the brain and about the visual system, there would still be something Mary would not know that her cohort with a more regular childhood would, namely, the nature of the experience of seeing a red tomato. Granted, she knows all about the neural states at work when someone sees a red tomato—after all, she has the utopian neuroscience at hand. What she would not know is what it is like to see red—what it is like to have that specific experience. Conclusion: her utopian neuroscience leaves something out. This omission implies that there is something in psychology that is not captured by neuroscience, which in turn implies that psychology cannot be reduced to neuroscience.

More formally and with some simplifications, the argument is this:

(F)

1. Mary knows everything there is to know about brain states and their properties.
2. It is not the case that Mary knows everything there is to know about sensations and their properties.
Therefore:

The argument is very interesting, and it gives an unusually clean line to the intuition that mental states are essentially private and have an irreducibly phenomenological character. Nonetheless I am not convinced, and I shall try to explain why.

First, I suspect that the intentional fallacy, which caused problems for arguments (A) and (B), likewise haunts the premises of argument (F). That aside, there are perhaps more revealing criticisms to be made. Paul M. Churchland (1985) and David Lewis (1983) have independently argued that “knows about” is used in different senses in the two premises. As they see it, one sense involves the manipulation of concepts, as when one knows about electromagnetic radiation and can use the concept “electromagnetic radiation” by having been tutored in the theory. The other sense involves a prelinguistic apprehension, as when one knows about electromagnetic radiation by having had one’s retina stimulated in the light of day, though one cannot use the expression “electromagnetic radiation.” The latter sense may involve innate dispositions to make certain discriminations, for example. If the first premise uses “knows about” in the first sense and the second uses it in the second sense, then the argument founders on the fallacy of equivocation.

The important point is this: if there are two (at least) modes of knowing about the world, then it is entirely possible that what one knows about via one method is identical to what one knows about via a different method. Pregnancy is something one can know about by acquiring the relevant theory from a medical text or by being pregnant. What a childless obstetrician knows about is the very same process as the process known by a pregnant but untutored woman. They both know about pregnancy. By parity of reasoning, the object of Mary’s knowledge when she knows the neurophysiology of seeing red might well be the very same state as the state known by her tomato-picking cohort. Just as the obstetrician does not become pregnant by knowing all about pregnancy, so Mary does not have the sensation of redness by knowing all about the neurophysiology of perceiving and experiencing red. Clearly it is no argument in support of nonidentity to say that Mary’s knowledge fails to cause the sensation of redness. Whichever suppose that it should.

There is a further reservation about this argument. With the first premise I take no issue, since we are asked to adopt it simply for the sake of argument. The second premise, in contrast, is supposed to be accepted because it is highly credible or perhaps dead obvious. Now although it does have a first blush plausibility, it is the premise on which the argument stands or falls, and closer scrutiny is required.

On a second look, its obviousness dissolves into contentiousness, because the premise asks me to be confident about something that is too far beyond the limits of what I know and understand. How can I assess what Mary will know and understand if she knows everything there is to know about the brain? Everything is a lot, and it means, in all likelihood, that Mary has a radically different and deeper understanding of the brain than anything barely conceivable in our wildest flights of fancy.

One might say well, if Mary knew everything about existing neuroscience, she would not know what it was like to experience red, and knowing absolutely everything will just be more of the same. That is an assumption to which the property dualist is not entitled to help himself. For to know everything about the brain might well be qualitatively different, and it might be to possess a theory that would permit exactly what the premise says it will not. First, utopian neuroscience will probably look as much like existing neuroscience as modern physics looks like Aristotelian physics. So it will not be just
more of the same. Second, all one need imagine is that Mary internalizes the theory in the way an engineer has internalized Newtonian physics, and she routinely makes introspective judgments about her own states using its concepts and principles. Like the engineer who does not have to make an effort but "sees" the world in a Newtonian manner, we may consider that Mary "sees" her internal world via the utopian neuroscience. Such a neuroscience might even tell her how to be very efficient at internalizing theories. It is, after all, the premise tells us, a complete neuroscience.

Intuitions and imaginability are, notoriously, a function of what we believe, and when we are very ignorant, our intuitions will be correspondingly naïve. Gedanken-experiments are the stuff of theoretical science, but when their venue is so surpassing distant from established science that the pivotal intuition is not uncontroversially better than its opposite, then their utility in deciding issues is questionable.

Moreover, intuitions opposite to those funding premise (2) are not only readily available, they can even be fleshe out a bit. How can I be reasonably sure that Mary would not know what a red tomato looks like? Here is a test. Present her with her first red object, and see whether she can recognize it as a red object. Given that she is supposed to know absolutely everything there is to know about the nervous system, perhaps she could, by introspective use of her utopian neuroscience, tell that she has, say, a gamma state in her O patterns, which she knows from her utopian neuroscience is identical to having a red sensation. Thus, she might recognize redness on that basis.

The telling point is this: whether or not she can recognize redness is clearly an empirical question, and I do not see how in our ignorance we can confidently insist that she must fail. Short of begging the question, there is no a priori reason why this is impossible. For all I know, she might even be able to produce red in her imagination if she knows what brain states are relevant. One cannot be confident that such an exercise of the imagination must be empirically impossible. To insist that our make-believe Mary could not make introspective judgments using her neuroscience because mental qualia are not identical to brain states would, obviously, route the argument round in a circle.

How could an alchemist assess what he could and could not know if he knew everything about substances? How could a monk living in the Middle Ages assess what he could and could not know if he knew everything there was to know about biology? He might insist, for example, that even if you knew everything there was to know about biology, you still would not know the nature of the vital spirit. Well, we still do not have a complete biology, but even so we know more than this hypothetical monk thought we could. We know (a) that there is no such thing as vital spirit, and (b) that DNA is the "secret" of life—it is what all living things on the planet share.

The central point of this reply to Jackson has been that he needs independent evidence for premise (2), since it is palpably not self-evident. It cannot be defended on a priori grounds, since its truth is an empirical question, and it cannot be defended on empirical grounds, since given the data so far, as good a case can be made for the negation of premise (2) as for premise (2) itself. I do not see, therefore, how it can be defended.

Concluding Remarks
There is a tendency to suppose that a dualist thesis, of either the substance or the property stripe, can be defended simply by noting and acknowledging that humans have an introspective capacity, that we are conscious, and that we are aware of our experiences. But it is a mistake to think that a reductionist need deny these observations. On the contrary, he might envision a reduction in which experiences, awareness, beliefs, perceptions, feelings, and so forth, are characterized in folk psychology, are identified with specific brain states. Such a reduction would fall on the retentive end of the inter-theoretic reduction spectrum, and hence it is a reduction that explicitly acknowledges the existence of mental states and consciousness, as now conceived. Nobody said light did not exist after the reduction of optics to electromagnetic theory; rather, they said light is electromagnetic radiation. And nobody said the laws of optics were useless or in disgrace. Notice that on this scenario, it would make perfectly good sense to talk about mental states causing brain states, since mental states turn out to be states of the brain (cf. Sperry 1980, Eccles 1977). Nor is a special notion of causation needed, as Sperry (1980) suggests.

Alternatively, a reductionist might envision a reduction in which a considerably evolved scientific psychology reduces to a considerably evolved neuroscience, such that our current conceptions of mental states and their nature are revised to some considerable extent. On this scenario, mental phenomena as we now conceive of them are not identified with brain states as we now conceive of them. This compares roughly with the bumpy reduction of Aristotelian physics to Newtonian physics, where "impetus" is not identified with anything in Newton’s theory, because the phenomenon is reconceived. Even here, the reductionist in no way denies the existence of mental phenomena. Rather, he expects that we may now misconceive and mis-
understand those phenomena and that a better, richer, and more comprehensive theory will be forthcoming. In the meantime he acknowledges the propriety of talking about mental phenomena in terms of whatever psychological theory is available, and he will enthusiastically support psychological research as a necessary part of an eventual reduction. (See chapter 9.) In the same way a molecular geneticist will enthusiastically support research in transmission genetics as a necessary part of an eventual reduction.

It is empirically possible that even a much-evolved and transformed psychology will fail to reduce to a much-evolved neuroscience. In that case mental phenomena, even newly conceived, will be emergent with respect to neural phenomena. Can we tell now that this will be the turn of the wheel? I think it is far too early to predict that psychology is forever beyond the reductive aspirations of neuroscience, and, as remarked earlier, one certainly cannot tell simply by observing the phenomena that no intertheoretic reduction is in the cards.

Nevertheless, there remains to be considered a set of powerful arguments that defend the hypothesis that psychology will not reduce to neuroscience. These arguments, still within the framework of property dualism, focus not on the phenomenological character of experience, not on the subjective point of view, but on the fact that mental states have a semantic dimension and that mental processes are, in some degree, logical. These arguments are the focus of the next section.