Chapter 3

Reduction, Qualia, and the Direct Introspection of Brain States

---

Do the phenomenological or qualitative features of our sensations constitute a permanent barrier to the reductive aspirations of any materialistic neuroscience? I here argue that they do not. Specifically, I wish to address the recent antireductionist arguments posed by Thomas Nagel (1974), Frank Jackson (1982), and Howard Robinson (1982). And I wish to explore the possibility of human subjective consciousness within a conceptual environment constituted by a matured and successful neuroscience.

If we are to deal sensibly with the issues here at stake, we must approach them with a general theory of scientific reduction already in hand, a theory motivated by and adequate to the many instances and varieties of interconceptual reduction displayed *elsewhere* in our scientific history. With an independently grounded account of the nature and grounds of intertheoretic reduction, we can approach the specific case of subjective qualia free from the myopia that results from trying to divine the proper conditions on reduction by simply staring long and hard at the problematic case at issue.

1 *Intertheoretic Reduction*

We may begin by remarking that the classical account of intertheoretic reduction (Nagel 1961) now appears to be importantly mistaken, though the repairs necessary are quickly and cleanly made. Suppressing nicities, we may state the original account as follows. A new and more comprehensive theory *reduces* an older theory just in case the new theory, when conjoined with appropriate correspondence rules, logically entails the principles of the older theory. (The point of the correspondence rules or "bridge laws" is to connect the disparate ontologies of the two theories; often these are expressed as identity statements, such as *Temperature* = $mv^2/3k$.) Schematically,

This essay first appeared in the *Journal of Philosophy* 82, no. 1 (January 1985).

T_N & (correspondence rules)

logically entails

T_O.

Difficulties with this view begin with the observation that most reduced theories turn out to be, strictly speaking and in a variety of respects, *false*. (Real gases don't really obey $PV = \mu RT$, as in classical thermodynamics; the planets don't really move in ellipses, as in Keplerian astronomy; the acceleration of falling bodies isn't really uniform, as in Galilean dynamics; etc.) If reduction is *de*duction, *modus tollens* would thus require that the premises of the new reducing theories (statistical thermodynamics in the first case, Newtonian dynamics in the second and third) be somehow false as well, in contradiction to their assumed truth.

This complaint can be temporarily deflected by pointing out that the premises of a reduction must often include, not just the new reducing theory, but also some limiting assumptions or counterfactual boundary conditions (such as that the molecules of a gas have only mechanical energy, or that the mass of the planets is negligible compared to the sun's, or that the distance any body falls is negligibly different from zero). Falsity in the reducing premises can thus be conceded, since it is safely confined to those limiting or counterfactual assumptions.

This defense will not deal with all cases of falsity, however, since in some cases the reduced theory is so radically false that some or all of its ontology must be rejected entirely, and the "correspondence rules" connecting that ontology to the newer ontology therefore display a problematic status. Newly conceived features cannot be identical with, nor even nomically connected with, old features, if the old features are illusory and uninstantiated. For example, relativistic mass is not identical with Newtonian mass, nor even coextensive with it, even at low velocities. Nevertheless, the reduction of Newtonian by Einsteinian mechanics is a paradigm of a successful reduction. For a second example, neither is caloric-fluid-pressure identical with, nor even coextensive with, mean molecular kinetic energy. But an overtly *fluid* thermodynamics (i.e., one committed to the existence of "caloric") still finds a moderately impressive reduction within statistical thermodynamics. In sum, even theories with a *nonexistent* ontology can enjoy reduction, and this fact is problematic on the traditional account at issue.

Cases like these invite us to give up the idea that what gets deduced in a reduction is the theory to be reduced. A more accurate,

general, and illuminating schema for intertheoretic reduction is as follows:

T_N & limiting assump. & boundary cond.

logically entails

I_N (a set of theorems of (restricted) T_N),
e.g., $(x)(Ax \supset Bx)$,
$(x)((Bx \,\&\, Cx) \supset Dx)$,

which is relevantly isomorphic with

T_O (the older theory),
e.g., $(x)(Jx \supset Kx)$,
$(x)((Kx \,\&\, Lx) \supset Mx)$.

That is to say, a reduction consists in the deduction, within T_N, not of T_O itself, but rather of a roughly equipotent *image* of T_O, an image still expressed in the vocabulary proper to T_N. The correspondence rules play no part whatever in the *de*duction. They show up only later, and not necessarily as material-mode statements, but as mere ordered pairs: $\langle Ax, Jx\rangle$, $\langle Bx, Kx\rangle$, $\langle Cx, Lx\rangle$, $\langle Dx, Mx\rangle$. Their function is to indicate which term substitutions in the image I_N will yield the principles of T_O. The older theory, accordingly, is never deduced; it is just the target of a relevantly adequate *mimicry*. Construed in this way, a correspondence rule is entirely consistent with the assumption that the older predicate it encompasses has no extension whatever. This allows that a true theory might reduce even a substantially false one.

The point of a reduction, according to this view, is to show that the new or more comprehensive theory contains explanatory and predictive resources which parallel, to a relevant degree of exactness, the explanatory and predictive resources of the reduced theory. The int*ra*-theoretic deduction (of I_N within T_N) and the int*er*theoretic mapping (of T_O into I_N) jointly constitute a fell-swoop demonstration that the older theory can be displaced wholesale by the new without significant explanatory or predictive loss. (This sketch of intertheoretic reduction is drawn from P. M. Churchland 1979, section 11. For a more detailed account, see Hooker 1981.)

Material-mode statements of identity can occasionally be made, of course. We do wish to assert that visible light = electromagnetic waves between 0.35 μm and 0.75 μm, that sound = atmospheric compression waves, that temperature = mean molecular kinetic energy, and that electric current = net motion of charged particles. But a correspondence rule does not itself make such a claim. At best, it records the fact that the new predicate applies in all those cases where its

T_O-doppelgänger predicate was normally *thought* to apply. On this view, full-fledged *identity* statements are licensed by the comparative *smoothness* of the relevant reduction (i.e., the limiting assumptions or boundary conditions on T_N are not wildly counterfactual, all or most of T_O's principles find close analogues in I_N, etc.). This smoothness permits the comfortable assimilation of the old ontology within the new, and thus allows the old theory to retain all or most of its ontological integrity. *It is smooth intertheoretic reductions that motivate and sustain statements of cross-theoretic identity, not the other way around.*

The preceding framework allows us to frame a useful conception of reduction for specific *properties*, as opposed to entire theories, and it allows us to frame a useful conception of the contrary notion of "emergent" properties. A property F, postulated by an older theory or conceptual framework T_O, is reduced to a property G in some new theory T_N just in case

(1) T_N reduces T_O,

(2) 'F' and 'G' are correspondence-rule paired in the reduction, and

(3) the reduction is sufficiently smooth to sustain the ontology of T_O and thus to sustain the identity claim 'F-ness = G-ness'.

Intuitively, and in the material mode, this means that F-ness reduces to G-ness just in case the causal powers of F-ness (as outlined in the laws of T_O) are a subset of the causal powers of G-ness (as outlined in the laws of T_N).

Finally, a property F will be said to be an *emergent* property (relative to T_N) just in case

(1) F is definitely real and instantiated,

(2) F is cooccurrent with some feature or complex circumstance recognized in T_N, but

(3) F cannot be *reduced* to any property postulated by or definable within T_N.

Intuitively, this will happen when T_N does not have the resources adequate to define a property with all of the causal powers possessed by F-ness. Claims about the emergence of certain properties are therefore claims about the relative poverty in the resources of certain aspirant theories. Having outlined these notions, we shall turn to address substantive questions of emergence and irreducibility in a few moments.

(A word of caution is perhaps in order here, since the expression 'emergent property' is often used in two diametrically opposed senses. In scientific contexts, one frequently hears it used to apply to what might be called a "network property," a property that appears exactly when the elements of some substrate are suitably organized, a property that *consists in* the elements of that substrate standing in certain relations to one another, a set of relations that collectively sustain the set of causal powers ascribed to the "emergent" property. In this innocent sense of 'emergent', there are a great many emergent properties, and quite probably the qualia of our sensations should be numbered among them. But in philosophical contexts one more often encounters a different sense of 'emergent', one that implies that an emergent property does *not* consist in any collective or organizational feature of its substrate. The first sense positively implies reducibility; the second implies *ir*reducibility. It is emergence in the second sense that is at issue in this paper.)

Before we continue, several points about reduction need to be emphasized. The first is that in arguing for the emergence of a given property F relative to some theory T_N, it is not sufficient to point out that the existence or appearance of F-ness cannot be deduced from T_N. It is occasionally claimed, for example, that the objective features of warmth or blueness must be irreducibly emergent properties, since however much one bends and squeezes the molecular theory concerning H_2O, one cannot deduce from it that water will be *blue*, but only that water will scatter electromagnetic radiation at such and such wavelengths. And however much one wrings from the mechanics of molecular motion, one cannot deduce from it that a roaring hearth will be *warm*, but only that its molecules will have such and such a mean kinetic energy and will collectively emit electromagnetic radiation at longish wavelengths.

These premises about nondeducibility are entirely true, but the conclusion against reducibility does not follow. It is a serious mistake to make even *in*direct deducibility (i.e., deducibility with the help of correspondence rules) a requirement on successful reduction, as we saw at the beginning of this section. And there are additional reasons why it would be even more foolish to insist on the much stronger condition of direct deducibility. For example, formal considerations alone guarantee that, for any predicate 'F' not already in the proprietary lexicon of the aspirant reducing theory T_N, no statements whatever involving 'F' (beyond tautologies and other trivial exceptions) will be deducible from T_N. The deducibility requirement would thus trivialize the notion of reduction by making it impossible for *any* conceptual framework to reduce any other, distinct conceptual framework. Even temperature, that paradigm of a successfully

reduced property, would be rendered irreducible, since the term 'temperature' does not appear in the lexicon of statistical mechanics.

There is a further reason why the demand for direct deducibility is too strong. The fact is, it is an historical accident that we humans currently use precisely the conceptual framework we do use. We might have used any one of an infinite number of other conceptual frameworks to describe the observable world, each one of which could have been roughly adequate to common experience, and many of which would be roughly isomorphic (each in its different way) with some part of the correct account that a utopian theory will eventually provide. Accordingly, we can legitimately ask of a putatively correct theory of a given objective domain that it account for the phenomena in (that is, function successfully in) that domain. But we cannot insist that it also be able to predict how this, that, or the other conceptually idiosyncratic human culture is going to *conceive* of that domain. That would be to insist that the new theory do *predictive cultural anthropology* for us, as well as mechanics, or electromagnetic theory, or what have you. The demand that molecular theory directly entail *our* thermal or color concepts is evidently this same unreasonable demand.

All we can properly ask of a reducing theory is that it have the resources to conjure up a set of properties whose nomological powers/roles/features are systematic *analogues* of the powers/roles/features of the set of properties postulated by the old theory. Since both theories presume to describe the same empirical domain, these systematic nomological parallels constitute the best grounds there can be for concluding that both theories have managed to latch onto the *same* set of objective properties. The hypothesized identity of the properties at issue explains why I_N and T_O are taxonomically and nomically parallel: they are both at least partially correct accounts of the very same objective properties. I_N merely frames that account within a much more penetrating conceptual system—that of T_N.

Moreover, it is to be expected that existing conceptual frameworks will eventually be reduced or displaced by new and better ones, and those in turn by frameworks better still, for who will be so brash as to assert that the feeble conceptual achievements of our adolescent species comprise an exhaustive account of anything at all? If we put aside this conceit, then the only alternatives to intertheoretic reduction are epistemic stagnation or the outright elimination of old frameworks as wholly false and illusory.

2 *Theoretical Change and Perceptual Change*

Esoteric properties and arcane theoretical frameworks are not the only things that occasionally enjoy intertheoretic reduction. Observ-

able properties and commonsense conceptual frameworks can also enjoy smooth reduction. Thus, being a middle-A sound is identical with being an oscillation in air pressure at 440 hertz; being red is identical with having a certain triplet of electromagnetic reflectance efficiencies; being warm is identical with having a certain mean level of microscopically embodied energies, and so forth.

Moreover, the relevant reducing theory is capable of replacing the old framework not just in contexts of calculation and inference. *It should be appreciated that the reducing theory can displace the old framework in all of its observational contexts as well.* Given the reality of the property identities just listed, it is quite open to us to begin framing our spontaneous perceptual reports in the language of the more sophisticated reducing theory. It is even desirable that we begin doing this, since the new vocabulary observes distinctions which are in fact within the discriminatory reach of our native perceptual systems, though those objective distinctions go unmarked and unnoticed from within the old framework. We can thus make more penetrating use of our native perceptual equipment. Such displacement is also desirable for a second reason: the greater inferential or computational power of the new conceptual framework. We can thus make better inferential *use* of our new perceptual judgments than we made of our old ones.

It is difficult to convey in words the enormity of such perceptual transformations and the naturalness of the new conceptual regime once established. A nonscientific example may help to get the initial point across.

Consider the enormous increase in discriminatory skill that spans the gap between an untrained child's auditory apprehension of a symphony, and the same person's apprehension of the same symphony forty years later, when hearing it in his capacity as conductor of the orchestra performing it. What was before a seamless voice is now a mosaic of distinguishable elements. What was before a dimly apprehended tune is now a rationally structured sequence of distinguishable and identifiable chords supporting an appropriately related melody line. The matured musician hears an entire world of structured detail, concerning which the child is both dumb and deaf.

Other modalities provide comparable examples. Consider the practiced and chemically sophisticated wine taster, for whom the category "red wine" used by most of us divides into a network of fifteen or twenty distinguishable elements: ethanol, glycol, fructose, sucrose, tannin, acid, carbon dioxide, and so forth, whose relative concentrations he can estimate with accuracy. Or consider the astronomer, for whom the speckled black dome of her youth has become a visible abyss, scattering nearby planets, yellow dwarf stars, blue and red

giants, distant globular clusters, and even a remote galaxy or two, all discriminable as such and locatable in three-dimensional space with her unaided (repeat: *unaided*) eye.

In each of these cases, what is finally mastered is a conceptual framework—whether musical, chemical, or astronomical—a framework that embodies far more wisdom about the relevant sensory domain than is immediately apparent to untutored discrimination. Such frameworks are characteristically a cultural heritage, pieced together over many generations, and their mastery supplies a richness and penetration to our sensory lives that would be impossible in their absence. (The role of theory in perception and the systematic enhancement of perception through theoretical progress are examined at length in P. M. Churchland 1979, sections 1 through 6.)

Our *introspective* lives are already the extensive beneficiaries of this phenomenon. The introspective discriminations we make are for the most part learned; they are acquired with practice and experience, often quite slowly. And the specific discriminations we learn to make are those it is useful for us to make. Generally, those are the discriminations that others are already making, the discriminations embodied in the psychological vocabulary of the language we learn. The conceptual framework for psychological states that is embedded in ordinary language is a modestly sophisticated theoretical achievement in its own right, and it shapes our matured introspection profoundly. If it embodied substantially *less* wisdom in its categories and connecting generalizations, our introspective apprehension of our internal states and activities would be much diminished, though our native discriminatory mechanisms remain the same. Correlatively, if folk psychology embodied substantially *more* wisdom about our inner nature than it actually does, our introspective discrimination and recognition could be very much *greater* than it is, though our native discriminatory mechanisms remain unchanged.

This brings me to the central positive suggestion of this paper. Consider now the possibility of learning to describe, conceive, and introspectively apprehend the teeming intricacies of our inner lives within the conceptual framework of a matured neuroscience, a neuroscience that successfully reduces, either smoothly or roughly, our commonsense folk psychology. Suppose we trained our native mechanisms to make a new and more detailed set of discriminations, a set that corresponded not to the primitive psychological taxonomy of ordinary language, but to some more penetrating taxonomy of states drawn from a completed neuroscience. And suppose we trained ourselves to respond to that reconfigured discriminative activity with judgments that were framed, as a matter of course, in the appropriate concepts

from neuroscience. (I believe it was Paul K. Feyerabend and Richard Rorty who first identified and explored this suggestion. See Feyerabend 1963a and Rorty 1965. This occurred in a theoretical environment prepared largely by Sellars 1956. The idea has been explored more recently in P. M. Churchland 1979 and in chapter 1 above.)

If the examples of the symphony conductor (who can hear the A*m7* chords), the enologist (who can see and taste the glycol), and the astronomer (who can see the temperature of a blue giant star) provide a fair parallel, then the enhancement in our introspective vision could approximate a revelation. Dopamine levels in the limbic system, the spiking frequencies in specific neural pathways, resonances in the *n*th layer of the occipital cortex, inhibitory feedback to the lateral geniculate nucleus, and countless other neurophysical nicities could be moved into the objective focus of our introspective discrimination, just as G*m7* chords and A*dim* chords are moved into the objective focus of a trained musician's auditory discrimination. We will of course have to *learn* the conceptual framework of a matured neuroscience in order to pull this off. And we will have to *practice* its noninferential application. But that seems a small price to pay for the quantum leap in self-apprehension.

All of this suggests that there is no problem at all in conceiving the eventual reduction of mental states and properties to neurophysiological states and properties. A matured and successful neuroscience need only include, or prove able to define, a taxonomy of kinds with a set of embedding laws that faithfully mimics the taxonomy and causal generalizations of *folk* psychology. Whether future neuroscientific theories will prove able to do this is a wholly empirical question, not to be settled a priori. The evidence for a positive answer is substantial and familiar, and it centers on the growing explanatory success of the several neurosciences.

But there is negative evidence as well; I have even urged some of it myself (1981a). My negative arguments there center on the explanatory and predictive poverty of folk psychology, and they question whether it has the categorial integrity to *merit* the reductive preservation of its familiar ontology. That line suggests substantial revision or outright elimination as the eventual fate of our mentalistic ontology. The qualia-based arguments of Nagel, Jackson, and Robinson, however, take a quite different line. They find no fault with folk psychology. Their concern is with the explanatory and descriptive poverty of any possible *neuroscience*, and their line suggests that emergence is the correct story for our mentalistic ontology. Let us now examine their arguments.

3 *Thomas Nagel's Arguments*

For Thomas Nagel, it is the phenomenological features of our experiences, the properties or *qualia* displayed by our sensations, that constitute a problem for the reductive aspirations of any materialistic neuroscience. In his classic position paper (1974), I find three distinct arguments in support of the view that such properties will never find any plausible or adequate reduction within the framework of a matured neuroscience. All three arguments are beguiling, but all three, I shall argue, are unsound.

The first argument

What makes the proposed reduction of mental phenomena different from reductions elsewhere in science, says Nagel, is that

> it is impossible to exclude the phenomenological features of experience from a reduction, in the same way that one excludes the phenomenal features of an ordinary substance from a physical or chemical reduction of it—namely, by explaining them as effects on the minds of human observers. (1974, p. 437)

The reason it is impossible to exclude them, continues Nagel, is that the phenomenological features are essential to experience, and to the subjective point of view. But this is not what interests me about this argument. What interests me is the claim that reductions of various substances elsewhere in science *exclude the phenomenal features of the substance.*

This is simply false, and the point is extremely important. The phenomenal features at issue are those such as the objective redness of an apple, the warmth of a coffee cup, and the pitch of a sound. These properties are not excluded from our reductions. Redness, an objective phenomenal property of apples, is identical with a certain wavelength triplet of electromagnetic reflectance efficiencies. Warmth, an objective phenomenal property of objects, is identical with the mean level of the objects' microscopically embodied energies. Pitch, an objective phenomenal property of a sound, is identical with its oscillatory frequency. These electromagnetic and micromechanical properties, out there in the objective world, are genuine phenomenal properties. Despite widespread ignorance of their dynamical and microphysical details, it is these objective physical properties to which everyone's perceptual mechanisms are keyed.

The reductions whose existence Nagel denies are in fact so complete that one can already displace entirely large chunks of our commonsense vocabulary for observable properties, and learn to frame

one's perceptual judgements directly in terms of the reducing theory. The mean kinetic energy (KE) of the molecules in this room, for example, is currently about 6.2×10^{-21} joules. The oscillatory frequency of this sound (I here whistle C one octave above middle C) is about 524 hertz. And the three critical electromagnetic reflectance efficiencies (at 0.45, 0.53, and 0.63 μm) of this white piece of paper are all above 80 percent. These microphysical and electromagnetic properties can be felt, heard, and seen, respectively. Our native sensory mechanisms can easily discriminate such properties, one from another, and their presence from their absence. They have been doing so for millennia. The "resolution" of these mechanisms is inadequate, of course, to reveal the microphysical details and the extended causal roles of the properties thus discriminated. But they are abundantly adequate to permit the reliable discrimination of the properties at issue. (See my 1979, sections 2 through 6. See also Paul and Patricia Churchland 1981b, pp. 128–130 [this volume, chapter 2, pp. 30–31].)

On this view, the standard perceptual properties are not "secondary" properties at all, in the standard sense that implies that they have no real existence save *inside* the minds of human observers. On the contrary, they are as objective as you please, with a wide variety of objective causal properties. Moreover, it would be a mistake even to *try* to "kick the phenomenal properties inwards," since that only postpones the problem of reckoning their place in nature. We shall only confront them again later as we address the place in nature of mental phenomena. And as Nagel correctly points out, the relocation dodge is no longer open to us once the problematic properties are already located within the mind.

Nagel concludes from this that subjective qualia are unique in being immune from the sort of reductions found elsewhere in science. I draw a very different conclusion. The *objective* qualia (redness, warmth, etc.) should never have been "kicked inwards to the minds of observers" in the first place. They should be confronted squarely, and they should be reduced where they stand: *out*side the human observer. As we saw, this can and has in fact been done. If objective phenomenal properties are so treated, then subjective qualia can be confronted with parallel forthrightness, and can be reduced where *they* stand: *in*side the human observer. So far then, the external and the internal cases are not different: they are parallel after all.

The second argument

A second argument urges the point that the intrinsic character of experiences, the qualia of sensations, are essentially accessible from

only a single point of view, the subjective point of view of the experiencing subject. The properties of physical brain states, by contrast, are accessible from a variety of entirely objective points of view. We cannot hope adequately to account for the former, therefore, in terms of properties appropriate to the latter domain. (see Nagel 1974, pp. 442–444.)

This somewhat diffuse argument appears to be an instance of the following argument.

(1) The qualia of my sensations are directly known by me, by introspection, as elements of my conscious self.

(2) The properties of my brain states are *not* directly known by me, by introspection, as elements of my conscious self.

∴ (3) The qualia of my sensations ≠ the properties of my brain states.

And perhaps there is a second argument here as well, a complement to the first:

(1) The properties of my brain states are known-by-the-various-external-senses, as having such and such physical properties.

(2) The qualia of my sensations are *not* known-by-the-various-external-senses, as having such and such physical properties.

∴ (3) The qualia of my sensations ≠ the properties of my brain states.

The argument form here is apparently

(1) Fa

(2) $\sim Fb$

∴ (3) $a \neq b$.

Given Leibniz's Law and the extensional nature of the property F, this is a valid argument form. But in the examples at issue, F is obviously not an extensional property. The fallacy committed in both cases is amply illustrated in the following parallel arguments.

(1) Hitler is widely recognized as a mass murderer.

(2) Adolf Schicklgruber is *not* widely recognized as a mass murderer.

∴ (3) Hitler ≠ Adolf Schicklgruber.

or,

(1) Aspirin is known by John to be a pain reliever.

(2) Acetylsalicylic acid is *not* known by John to be a pain reliever.

∴ (3) Aspirin ≠ acetylsalicylic acid.

or, to cite an example very close to the case at issue,

(1) Temperature is known by me, by tactile sensing, as a feature of material objects.

(2) Mean molecular kinetic energy is *not* known by me, by tactile sensing, as a feature of material objects.

∴ (3) Temperature ≠ mean molecular kinetic energy.

The problem with all of these arguments is that the "property" ascribed in premise (1) and witheld in premise (2) consists only in the subject item's being *recognized, perceived,* or *known* as something, *under some specific description or other.* Such apprehension is not a genuine feature of the item itself, fit for divining identities, since one and the same subject may be successfully recognized under one description (e.g., 'qualia of my mental state'), and yet fail to be recognized under another, equally accurate, coreferential description (e.g., 'property of my brain state'). In logician's terms, the propositional function, '*x* is known (perceived, recognized) by me, as an *F*' is one of a large number of *intensional contexts* whose distinguishing feature is that they do not always retain the same truth value through substitution of a coreferential or coextensive term for whatever holds the place of '*x*'. Accordingly, that such a context (i.e., the one at issue) should show a difference in truth value for two terms '*a*' and '*b*' (i.e., 'qualia of my sensations' and 'property of my brainstates') is therefore hardly grounds for concluding that '*a*' and '*b*' cannot be coreferential or coextensive terms! (I believe it was Richard Brandt and Jaegwon Kim (1967) who first identified this fallacy specifically in connection with the identity theory.)

This objection is decisive, I think, but it does not apply to a different version of the argument, which we must also consider. It may be urged that one's brain states are more than merely not (yet) known by introspection: they are not know*able* by introspection under any circumstances. In correspondence, Thomas Nagel has advised me that what he wishes to defend is the following *modalized* version of the argument.

(1) My mental states are knowable by me by introspection.

(2) My brain states are *not* knowable by me by introspection.

∴ (3) My mental states ≠ my brain states.

Here Nagel will insist that being knowable by me by introspection is a genuine relational property of a thing, and that this version of the argument is free of the intensional fallacy discussed above.

And so it is. But now the reductionist is in a position to insist that the argument contains a false premise: premise (2). At the very least, he can insist that (2) begs the question. For if mental states are indeed identical with brain states, then it is really brain states that we have been introspecting all along, though without appreciating their fine-grained nature. And if we can learn to think of and recognize those states under their familiar mentalistic descriptions—*as all of us have*—then we can certainly learn to think of and recognize them under their more penetrating neurophysiological descriptions. Brain states, that is, are indeed know*able* by introspection, and Nagel's argument commits the same error instanced below.

(1) Temperature is knowable by tactile sensing.

(2) Mean molecular kinetic energy is *not* knowable by tactile sensing.

∴ (3) Temperature ≠ mean molecular kinetic energy.

Here the conclusion is known to be false. Temperature is indeed mean molecular kinetic energy. Since the argument is valid, it must therefore have a false premise. Premise (2) is clearly the stinker. Just as one can learn to feel that the summer air is about 70°F, or 21°C, so one can learn to feel that the mean KE of its molecules is about 6.2×10^{-21} joules, for whether we realize it or not, that is the property our native discriminatory mechanisms are keyed to. And if one can come to know, by feeling, the mean KE of atmospheric molecules, why is it unthinkable that one might come to know, by introspection, the states of one's brain? (What would that feel like? It would feel exactly the same as introspecting the states of one's mind, since they are one and the same states. One would simply employ a different and more penetrating conceptual framework in their description.)

One must be careful, in evaluating the plausibility of Nagel's second premise, to distinguish it from the second premise of the very first version of the argument, the version that commits the intensional fallacy. My guess is that Nagel has profited somewhat from the ambiguity here. For in the first version, both premises are true. And in the second version, the argument is valid. Neither version, however, meets both conditions.

The matter of introspecting one's brain states will arise once more in the final section of this paper. For now, let us move on.

The third argument

The last argument here is the one most widely associated with Nagel's paper. The leading example is the (mooted) character of the experiences enjoyed by an alien creature such as a bat. The claim is that, no matter how much one knew about the bat's neurophysiology and its interaction with the physical world, one could still not know, nor perhaps even imagine, what it is like to be a bat. Even total knowledge of the physical details still leaves something out. The lesson drawn is that the reductive aspirations of neurophysiology are doomed to dash themselves, unrealized, against the impenetrable keep of subjective qualia. (see Nagel 1974, pp. 438ff.)

This argument is almost identical to an argument put forward in a recent paper by Frank Jackson (1982). Since Jackson's version deals directly with humans, I shall confront the problem as he formulates it.

4 *Jackson's Knowledge Argument*

Imagine a brilliant neuroscientist named Mary who has lived her entire life in a room that is rigorously controlled to display only various shades of black, white, and grey. She learns about the outside world by means of a black/white television monitor, and being brilliant, she manages to transcend these obstacles. She becomes the world's greatest neuroscientist, all from within this room. In particular, she comes to know everything there is to know about the physical structure and activity of the brain and its visual system, of its actual and possible states.

But there would still be something she did *not* know, and could not even imagine, about the actual experiences of all the other people who live outside her black/white room, and about her possible experiences were she finally to leave her room: the nature of the experience of seeing a ripe tomato, what it is like to see red or have a sensation-of-red. Therefore, complete knowledge of the physical facts of visual perception and its related brain activity *still leaves something out*. Hence, materialism cannot give an adequate reductionist account of all mental phenomena.

To give a conveniently tightened version of this argument,

(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, by Leibniz's Law,

(3) Sensations and their properties ≠ brain states and their properties

It is tempting to insist that we here confront just another instance of the intensional fallacy discussed earlier, but Jackson's defenders (e.g., Campbell 1983) insist that 'knows *about*' is a perfectly transparent, entirely extensional context. Let us suppose that it is. We can, I think, find at least two other shortcomings in this sort of argument.

The first shortcoming

This defect is simplicity itself. 'Knows about' may be transparent in both premises, but it is not *univocal* in both premises. (David Lewis [1983] and Laurence Nemirow [1980] have both raised this same objection, though their analysis of the ambiguity at issue differs from mine.) Jackson's argument is valid only if 'knows about' is univocal in both premises. But the kind of knowledge addressed in premise (1) seems pretty clearly to be different from the kind of knowledge addressed in (2). Knowledge in (1) seems to be a matter of having mastered a set of sentences or propositions, the kind one finds written in neuroscience texts; whereas knowledge in (2) seems to be a matter of having a representation of redness in some prelinguistic or sublinguistic medium of representation for sensory variables, or to be a matter of being able to *make* certain sensory discriminations, or something along these lines.

Lewis and Nemirow plump for the "ability" analysis of the relevant sense of 'knows about', but they need not be so narrowly committed, and the complaint of equivocation need not be so narrowly based. As my alternative gloss illustrates, other analyses of 'knowledge by acquaintance' are possible, and the charge of equivocation will be sustained so long as the type of knowledge invoked in premise (1) is distinct from the type invoked in premise (2). Importantly, they do seem very different, even in advance of a settled analysis of the latter.

In short, the difference between a person who knows all about the visual cortex but has never enjoyed a sensation of red, and a person who knows no neuroscience but knows well the sensation of red, may reside not in *what* is respectively known by each (brain states by the former, qualia by the latter), but rather in the different *type* of knowledge each has of *exactly the same thing*. The difference is in the

manner of the knowing, not in the nature of the thing(s) known. If one replaces the ambiguous occurrences of 'knows about' in Jackson's argument with the two different expansions suggested above, the resulting argument is a clear *non sequitur*.

(*a*) Mary has mastered the complete set of true propositions about people's brain states.

(*b*) Mary does *not* have a representation of redness in her prelinguistic medium of representation for sensory variables.

Therefore, by Leibniz's Law,

(*c*) The redness sensation $\neq$ any brain state.

Premises (*a*) and (*b*) are compossible, even on a materialist view. But they do not entail (*c*).

In sum, there are pretty clearly more ways of "having knowledge" than having mastered a set of sentences. And nothing in materialism precludes this. The materialist can freely admit that one has "knowledge" of one's sensations in a way that is independent of the scientific theories one has learned. This does not mean that sensations are beyond the reach of physical science. *It just means that the brain uses more modes and media of representation than the simple storage of sentences.* And this proposition is pretty obviously true: almost certainly the brain uses a considerable variety of modes and media of representation, perhaps hundreds of them. Jackson's argument, and Nagel's, exploit this variety illegitimately: both arguments equivocate on 'knows about'.

This criticism is supported by the observation that, if Jackson's form of argument were sound, it would prove far too much. Suppose that Jackson were arguing not against materialism, but against dualism: against the view that there exists a nonmaterial substance—call it 'ectoplasm'—whose hidden constitution and nomic intricacies ground all mental phenomena. Let our cloistered Mary be an "ectoplasmologist" this time, and let her know_1 (by description) everything there is to know about the ectoplasmic processes underlying vision. There would still be something she did not know_2 (by acquaintance): what it is like to see red. Dualism is therefore inadequate to account for all mental phenomena.

This argument is as plausible as Jackson's, and for the same reason: it exploits the same equivocation. But the truth is, such arguments show nothing, one way or the other, about how mental phenomena might be accounted for.

The second shortcoming

There is a further shortcoming with Jackson's argument, one of profound importance for understanding one of the most exciting consequences to be expected from a successful neuroscientific account of mind. I draw your attention to the assumption that even a utopian knowledge of neuroscience *must* leave Mary hopelessly in the dark about the subjective qualitative nature of sensations not yet enjoyed. It is true, of course, that no sentence of the form '*x* is a sensation-of-red' will be deducible from premises restricted to the language of neuroscience. But this is no point against the reducibility of phenomenological properties. As we saw in section 1, direct deducibility is an intolerably strong demand on reduction, and if this is all the objection comes to, then there is no objection worth addressing. What the defender of emergent qualia must have in mind here, I think, is the claim that Mary could not even *imagine* what the relevant experience would be like, despite her exhaustive neuroscientific knowledge, and hence, that she must still be missing certain crucial information.

This claim, however, is simply false. Given the truth of premise (1), premise (2) seems plausible to Jackson, Nagel, and Robinson only because none of these philosophers has adequately considered how much one might know if, as premise (1) asserts, one knew *everything* there is to know about the physical brain and nervous system. In particular, none of these philosophers has even begun to consider the changes in our introspective apprehension of our internal states that could follow upon a wholesale revision in our conceptual framework for our internal states.

The fact is, we can indeed imagine how neuroscientific information would give Mary detailed information about the qualia of various sensations. Recall our earlier discussion of the transformation of perception through the systematic reconceptualization of the relevant perceptual domain. In particular, suppose that Mary has learned to conceptualize her inner life, even in introspection, in terms of the completed neuroscience we are to imagine. So she does not identify her visual sensations crudely as 'a sensation-of-black', 'a sensation-of-grey', or 'a sensation-of-white'; but rather identifies them more revealingly as various spiking frequencies in the *n*th layer of the occipital cortex (or whatever). If Mary has the relevant neuroscientific concepts for the sensational states at issue (namely, sensations-of-*red*), but has never yet been *in* those states, she may well be able to imagine being in the relevant cortical state, and imagine it with substantial success, even in advance of receiving external stimuli that would actually produce it.

One test of her ability in this regard would be to give her a stimulus that would (finally) produce in her the relevant state (namely, a spiking frequency of 90 hertz in the gamma network: a "sensation-of-red" to us), and see if she can identify it correctly *on introspective grounds alone*, as 'a spiking frequency of 90 hertz, the kind a tomato would cause'. It does not seem to me to be impossible that she should succeed in this, and do so regularly on similar tests for other states, conceptualized clearly by her, but not previously enjoyed.

This may seem to some an outlandish suggestion, but the following will show that it is not. Musical chords are auditory phenomena that the young and unpracticed ear hears as undivided wholes, discriminable one from another, but without elements or internal structure. A musical education changes this, and one comes to hear chords as groups of discriminable notes. If one is sufficiently practised to have absolute pitch, one can even name the notes of an apprehended chord. And the reverse is also true: if a set of notes is specified verbally, a trained pianist or guitarist can identify the chord and recall its sound in auditory imagination. Moreover, a really skilled individual can construct, in auditory imagination, the sound of a chord he may never have heard before and certainly does not remember. Specify for him a relatively unusual one—an F♯9th*add*13th for example—and let him brood for a bit. Then play for him three or four chords, one of which is the target, and see if he can pick it out as the sound that meets the description. Skilled musicians can do this. Why is a similar skill beyond all possibility for Mary?

Ah, it is tempting to reply, musicians can do this only because chords are audibly structured sets of elements. Sensations-of-color are not.

But neither did chords seem, initially, to be structured sets of elements. They also seemed to be undifferentiated wholes. Why should it be unthinkable that sensations-of-color possess a comparable internal structure, unnoticed so far, but awaiting our determined and informed inspection? Jackson's argument, to be successful, must rule this possibility out, and it is difficult to see how he can do this *a priori*, especially since there has recently emerged excellent empirical evidence to suggest that *our sensations-of-color are indeed structured sets of elements*.

The Retinex theory of color vision recently proposed by Edwin Land (1977) represents any color apprehendable by the human visual system as being uniquely specified by its joint position along three vertices—its reflectance efficiencies at three critical wavelengths, those wavelengths to which the retina's triune cone system is selectively responsive. Since colors are apprehended by us, it is a good

hypothesis that those three parameters are represented in our visual systems, and that our sensations-of-color are in some direct way determined by them. Sensations-of-color may turn out literally to *be* three-element chords in some neural medium! In the face of all this, I do not see why it is even briefly plausible to insist that it is utterly impossible for a conceptually sophisticated Mary accurately to imagine, and then reliably pick out, color sensations she has not previously enjoyed. We can already foresee how it might actually be done.

The preceding argument does not collapse the distinction (between knowledge by description and knowledge by acquaintance) urged earlier in the discussion of equivocation. But it does show that the "taxonomies" that reside in our prelinguistic media of representation can be profoundly shaped by the taxonomies that reside in the linguistic medium, especially if one has had long practice at the observational discrimination of items that answer to those linguistically embodied categories. This is just a further illustration of the plasticity of human perception.

I do not mean to suggest, of course, that there will be no limits to what Mary can imagine. Her brain is finite, and its specific anatomy will have specific limitations. For example, if a bat's brain includes computational machinery that the human brain simply lacks (which seems likely), then the subjective character of *some* of the bat's internal states may well be beyond human imagination. Clearly, however, the elusiveness of the bat's inner life here stems not from the metaphysical "emergence" of its internal qualia, but only from the finite capacities of our idiosyncratically human brains. Within those sheerly structural limitations, our imaginations may soar far beyond what Jackson, Nagel, and Robinson suspect, if we possess a neuroscientific conceptual framework that is at last adequate to the intricate phenomena at issue. (See especially chapter 5, section 7, and chapter 9, section 4.)

I suggest, then, that those of us who prize the flux and content of our subjective phenomenological experience need not view the advance of materialistic neuroscience with fear and foreboding. Quite the contrary. The genuine arrival of a materialist kinematics and dynamics for psychological states and cognitive processes will constitute not a gloom in which our inner life is suppressed or eclipsed, but rather a dawning, in which its marvellous intricacies are finally *revealed*—most notably, if we apply ourselves, in direct self-conscious introspection.