Knowing Qualia: A Reply to Jackson

Paul M. Churchland (in *A Neurocomputational Perspective*, MIT Press, 1989)

In a recent paper concerning the direct introspection of brain states (1985b) I leveled three criticisms against Frank Jackson's "knowledge argument." At stake was his bold claim that no materialist account of mind can possibly account for all mental phenomena. Jackson has replied to those criticisms in his 1986. It is to those replies, and to the issues that prompted them, that the present chapter is directed.

1 *The Persistent Equivocation*

Jackson concedes the criticism I leveled at my own statement of his argument -- specifically, that it involves an equivocation on 'knows about' -- but he insists that my reconstruction does not represent the argument he wishes to defend. I accept his instruction, and turn my attention to the summary of the argument he provides at the bottom of page 293. Mary, you will recall, has been raised in innocence of any color experience, but has an exhaustive command of neuroscience.

(1) Mary (before her release) knows everything physical there is to know about other people.

(2) Mary (before her release) does not know everything there is to know about other people (because she *learns* something about them on her release).

(3) There are truths about other people (and herself) which escape the physicalist story.

Regimenting further, for clarity's sake, yields the following.[1]

(1) $(\forall x)[(Hx \ \& \ Px) \rightarrow Kmx]$

(2) $(\exists x)[Hx \ \& \ \sim Kmx]$ (viz., "what it is like to see red")

$\therefore$ (3) $(\exists x)[Hx \ \& \ \sim Px]$

Here *m* = Mary; *Kyx* = *y* knows about *x; Hx* = *x* is about persons; *Px* = *x* is about something physical in character; and *x* ranges over "knowables," generously construed so as not to beg any questions about whether they are propositional or otherwise in nature.

Thus expressed, the argument is formally valid: the salient move is a *modus tollens* that applies the second conjunct of premise (2), *'~Kmx',* to the waiting consequent of premise (1), *'Kmx'.* The questions now are whether the premises are jointly true, and whether the crucial notion *'Kmx'* is univocal in both of its appearances. Here lam surprised that Jackson sees any progress at all with the above formulation, since I continue to see the same equivocation found in my earlier casting of his argument.

Specifically, premise (1) is plausibly true, within Jackson's story about Mary's color-

[1] [Translation: (1) For all knowables x, if x is about humans and physical, then Mary knows x. (2) There is something x that is about humans but Mary does not know it. (3) Therefore, there is something x about humans that is not physical.]

free upbringing, only on the interpretation of 'knows about' that casts the object of knowledge as something propositional, as something adequately expressible in an English sentence. Mary, to put it briefly, gets 100 percent on every written and oral exam; she can pronounce on the truth of any given sentence about the physical characteristics of persons, especially the states of their brains. Her "knowledge by description" of physical facts about persons is without lacunae .

Premise (2), however, is plausibly true only on the interpretation of 'knows about' that casts the object of knowledge as something nonpropositional, as something inarticulable, as something that is non–truth-valuable. What Mary is missing is some form of "knowledge by acquaintance," acquaintance with a sensory character, prototype, or universal, perhaps.

Given this prima facie difference in the sense of 'knows about', or the kind of knowledge appearing in each premise, we are still looking at a prima facie case of an argument invalid by reason of equivocation on a critical term. Replace either of the 'K's above with a distinct letter, as acknowledgment of the ambiguity demands, and the inference to (3) evaporates. The burden of articulating some specific and unitary sense of 'knows about', and of arguing that both premises are true under that interpretation of the epistemic operator, is an undischarged burden that still belongs to Jackson.

It is also a *heavy* burden, since the resources of modern cognitive neurobiology already provide us with a plausible account of what the difference in the two kinds of knowledge amounts to, and of how it is possible to have the one kind without the other. Let me illustrate with a case distinct from that at issue, so as not to beg any questions.

Any competent golfer has a detailed representation (perhaps in his cerebellum, perhaps in his motor cortex) of a golf swing. It is a *motor* representation, and it constitutes his "knowing how" to *execute* a proper swing. The same golfer will also have a discursive representation of a golf swing (perhaps in his language cortex, or in the neighboring temporal and parietal regions), which allows him to describe a golf swing or perhaps draw it on paper. The motor and the discursive representations are quite distinct. Localized brain trauma, or surgery, could remove either one while sparing the other. Short of that, an inarticulate golf champion might have a superb representation of the former kind, but a feeble representation of the latter kind. And a physicist or sports physiologist might have a detailed and penetrating representation of the mechanics of a good swing, and yet be unable to duff the ball more than ten feet because he lacks an adequate *motor* representation, of the desired behavioral sequence, in the brain areas that control his limbs. Indeed, if our physicist is chronically disabled in his motor capacities, he may have no motor representation of a golf swing whatsoever. In one medium of representation, his representational achievements on the topic may be complete; while in another medium of representation, he has nothing.

A contrast between "knowing how" and "knowing that" is one already acknowledged in common sense, and thus it is not surprising that some of the earliest replies to Jackson's argument (Nemirow 1980; Lewis 1983) tried to portray its equivocation in these familiar terms, and tried to explicate Mary's missing knowledge solely in terms of her missing some one or more *abilities* (to recognize red, to imagine red, etc.). While the approach is well motivated, this binary distinction in types of knowledge barely begins to suggest the range and variety of different sites and types of internal representation to be found in a normal brain. There is no reason why we must be bound by the crude divisions of our pre scientific idioms when we attempt to give a precise and positive explication of

the equivocation displayed in Jackson's argument. And there are substantial grounds for telling a somewhat different story concerning the sort of nondiscursive knowledge at issue. Putting caution and qualification momentarily aside, I will tell such a story .

In creatures with trichromatic vision (i.e., with three types of retinal cone), color information is coded as a pattern of spiking frequencies across the axonal fibers of the parvocellular subsystem of the optic nerve. That massive cable of axons leads to a second population of cells in a central body called the lateral geniculate nucleus (LGN), whose axonal projections lead in turn to the several areas of the visual cortex at the rear of the brain's cerebral hemispheres, to VI, V2, and ultimately to V 4, which area appears to be especially devoted to the processing and representation of color information (Zeki 1980; Van Essen and Maunsell 1983; Hubel and Livingstone 1987). Human cognition divides a smooth continuum of color inputs into a finite number of prototypical categories. The laminar structure at V4 is perhaps the earliest place in the processing hierarchy to which we might ascribe that familiar taxonomy. A creature competent to make reliable color discriminations has there developed a representation of the range of familiar colors, a representation that appears to consist in a specific configuration of weighted synaptic connections meeting the millions of neurons that make up area V4.

That configuration of synaptic weights partitions the "activation space" of the neurons in area V4: it partitions that abstract space into a structured set of subspaces, one for each prototypical color. Inputs from the eye will each occasion a specific pattern of activity across these cortical neurons, a pattern or vector that falls within one of those subspaces. In such a pigeon holing, it now appears, does visual recognition of a color consist (see chapters 5 and 9 for the general theory of information processing here appealed to). This recognition depends upon the creature possessing a prior representation-a learned configuration of synapses meeting the relevant population of cells-that antecedently partitions the creature's visual taxonomy so it can respond selectively and appropriately to the flux of visual stimulation arriving from the retina and LGN .

This distributed representation is not remotely propositional or discursive, but it is entirely real. All trichromatic animals have one, even those without any linguistic capacity. It apparently makes possible the many abilities we expect from color-competent creatures: discrimination, recognition, imagination, and so on. Such a representation is presumably what a person with Mary's upbringing would lack, or possess only in stunted or incomplete form. Her representational space within the relevant area of neurons would contain only the subspace for black, white, and the intervening shades of gray, for the visual examples that have shaped her synaptic configuration were limited to these. There is thus more than just a clutch of abilities missing in Mary: there is a complex representation-a processing framework that deserves to be called "cognitive"-that she either lacks or has in reduced form. There is indeed something she "does not know." Jackson's premise (2), we may assume, is thus true on these wholly materialist assumptions.

These same assumptions are entirely consistent with the further assumption that elsewhere in Mary's brain-in the language areas, for example-she has stored a detailed and even exhaustive set of discursive, propositional, truth-valuable representations of what goes on in people's brains during the experience of color, a set she has brought into being by the exhaustive reading of authoritative texts in a completed cognitive neuroscience. She may even be able to explain her own representational deficit, as sketched above, in complete neurophysical detail. Jackson's premise (1), we may thus

assume, is also true on these wholly materialist assumptions.

The view sketched above is alive candidate for the correct story of sensory coding and sensory recognition. But whether or not it is true, it is at least a logical possibility. Accordingly, what we have sketched here is a consistent but entirely *physical* model (i.e., a model in which Jackson's conclusion is false) in which both of Jackson's premises are true under the appropriate interpretation. They can hardly entail a conclusion, then, that is inconsistent with physicalism. Their compossibility, on purely physicalist assumptions, resides in the different character and the numerically different medium of representation at issue in each of the two premises. Jackson's argument, to re file the charge, equivocates on 'knows about'.

2 *Other Invalid Instances*

An argument form with one invalid instance can be expected to have others. This was the point of a subsidiary objection in my 1985b paper: if valid, Jackson's argument, or one formally parallel, would also serve to refute the possibility of *substance dualism.* I did not there express my point with notable clarity, however, and I accept responsibility for Jackson's quite missing my intention. Let me try again.

The basic point is that the canonical presentation of the knowledge argument, as outlined on p. 67 above, would be just as valid if the predicate term *'P'* were everywhere replaced by *'E'.* And the resulting premises would be just as plausibly true if

(1) *'E'* stood for 'is about something ectoplasmic in character' (where 'ectoplasm' is an arbitrary name for the dualist's nonphysical substance), and

(2) the story is altered so that Mary becomes an exhaustive expert on a completed *ectoplasmic* science of human nature.

The plausibility would be comparable, I submit, because a long discursive lecture on the objective, statable, law-governed properties of ectoplasm, whatever they might be, would be exactly as useful, or *useless,* in helping Mary to *know-by-acquaintance* "what it is like to see red," as would along discursive lecture on the objective, statable, law-governed properties of the physical matter of the brain. Even if substance dualism were true, therefore, and ectoplasm were its heroic principal, an exactly parallel 'knowledge argument" would "show" that there are some aspects of consciousness that must forever escape the *ectoplasmic* story. Given Jackson's antiphysicalist intentions, it is at least an irony that the same form of argument should incidentally serve to blow substance dualism out of the water .

Though lam hardly a substance dualist (and neither is Jackson), I do regard substance dualism as a theoretical possibility, one that might conceivably succeed in explicating the psychological ontology of common sense in terms of the underlying properties and law-governed behavior of the nonmaterial substance it postulates. And I must protest that the parallel knowledge argument against substance dualism would be wildly unfair, and for the very same reason that its analogue against physicalism is unfair: it would equivocate on 'knows about'. It would be no more effective against dualism than it is against materialism.

The parallel under examination contains a further lesson. If it works at all, Jackson's argument works against physicalism not because of some defect that is unique to

physicalism; *it works because no amount of discursive knowledge, on any topic, will constitute the nondiscursive form of knowledge that Mary lacks.* Jackson's argument is one instance of an indiscriminately *antireductionist* form of argument. If it works at all, an analog will work against any proposed reductive, discursive, objective account of the nature of our subjective experience, no matter what the reducing theory might happen to be. I see this as a further symptom of the logical pathology described earlier . Since the argument "works" for reasons that have nothing essential to do with physicalism, it should "work" against the explanatory aspirations of other ontologies as well. And so it "does." The price of embracing Jackson's argument is thus dramatically higher than first appears. For it makes any scientific account of our sensory experience entirely impossible, no matter what the ontology employed.

3 *A Genuinely Nonequivocal Knowledge Argument*

We can appreciate the equivocation more deeply if we explore a version of Jackson's argument that does *not* equivocate on 'knows about'. The equivocation can quickly be closed, if we are determined to do so, and the results are revealing. Given that the problem is a variety in the possible forms of knowing, let us simply rewrite the argument with suitable quantification over the relevant forms of knowing. The first premise must assert that, for any knowable *x,* and for any form *f* of knowledge, if *x* is about humans and *x* is physical in character, then Mary *knows(f)* about *x.* The second premise is modified in the same modest fashion, and the conclusion is identical. Canonically,[2]

(1') $(\forall x)(\forall f)[(Hx \ \& \ Px) \rightarrow K(f)mx]$

(2') $(\exists x)(\exists f)[Hx \ \& \sim K(f)mx]$

$\therefore$ (3') $(\exists x)[Hx \ \& \sim Px]$

This argument is also formally valid, and its premises explicitly encompass whatever variety there may be in forms of knowing. What can we say about its soundness?

Assume that Mary has had the upbringing described in Jackson's story , and thus lacks any knowledge-by-acquaintance with "what it is like to see red." Premise (2') will then be true, as and for the reasons that Jackson's story requires. What will be the truth value of premise (1') on these assumptions?

Premise (1') is now a very strong claim indeed, much stronger than the old premise (1), and a materialist will be sure to insist that it is false. The reason offered will be that, because of her deprived upbringing, Mary quite clearly *lacks* one form of knowledge of a certain physical aspect of people. Specifically, she lacks a proper configuration of synaptic connections meeting the neurons in the appropriate area of her visual cortex. She thus lacks an appropriately partitioned activation vector space across those neurons, and therefore has no representation, at that site, of the full range of sensory coding vectors that might someday come from the retina and the LGN. In other words, there is something physical about persons (their color sensations, or identically, their coding

[2] [Translation: (1) For all knowables x and forms of knowledge f, if x is about humans and physical, then Mary knows x in the f way. (2) There is a knowable x and a form of knowledge f such that x is about humans and Mary does not know it in the f way. (3) Therefore, there is something x that is about humans that is not physical.]

vectors in their visual pathways), and there is some form of knowledge (an antecedently partitioned prelinguistic taxonomy), such that Mary lacks that form of knowledge of that aspect of persons. Accordingly, premise (1') is false and the conclusion (3') is not sustained.

From a materialist's point of view, it is obvious that (1') will be false on the assumptions of Jackson's story. For that story denies her the upbringing that normally provokes and shapes the development of the relevant representation across the appropriate population of cortical neurons. And so, of course, there is a form of knowledge, of a physical aspect of persons, that Mary does not have. As just illustrated, the materialist can even specify that form of knowledge, and its objects, in neural terms. But this means that premise (l'), as properly quantified at last, is false. Mary does *not* have knowledge of everything physical about persons, in every way that is possible for her. (That is why premise (2') is true.)

There is, of course, no guarantee that the materialist's account of sensations and sensory recognition is correct (although the experimental and theoretical evidence for a view of this general kind continues to accumulate). But neither is Jackson in a position to insist that it must be mistaken. That would beg the very question at issue: whether sensory qualia form a metaphysically distinct class of phenomena beyond the scope of physical science.

To summarize, if we write a deliberately non-equivocal form of Jackson's argument, one that quantifies appropriately over all of the relevant forms of knowledge, then the first premise must almost certainly be false under the conditions of his own story .So, at any rate, is the materialist in a strong position to argue. Jackson's expressed hope for highly plausible premises" is not realized in (1'). The original premise (1) was of course much more plausible. But it failed to sustain a valid argument, and it was plausible only because it failed to address all the relevant forms of knowledge.

4 *Converting a Third-Person Account into a First-Person Account*

My final objection to Jackson was aimed more at breaking the grip of the ideology behind his argument than at the argument itself. That ideology includes a domain of properties-the qualia of subjective experience-that are held to be metaphysically distinct from the objective physical properties addressed by orthodox science. It is not a surprise, then, on this view, that one might know all physical facts, and yet be ignorant of some domain of these nonphysical qualia. The contrast between what is known and what is not known simply reflects an antecedent metaphysical division in the furniture of the world.

But there is another way to look at the situation, one that finds no such division. Our capacity for recognizing a range of (currently) inarticulable features in our subjective experience is easily explained on I materialist principles; the relevant sketch appears earlier in this essay and elsewhere in this volume (chapter 5, section 7). Our discursive inarticulation of those features is no surprise either, and signifies nothing about their metaphysical status (chapter 10, section 5). Indeed, that veil of inarticulation may itself be swept aside by suitable learning. What we are now able spontaneously to report about our internal states and cognitive activities need not define the limit on what we might be able to report, spontaneously and accurately, if we were taught a more appropriate conceptual scheme in which to express our discriminations. In closing, let me again urge on Jackson this exciting possibility.

The intricacies of brain function may be subjectively opaque to us now, but they need not remain that way forever. Neuroscience may appear to be defective in providing a purely "third-person account" of mind, but only familiarity of idiom and spontaneity of conceptual response are required to make it a "first-person account" as well. *What makes an account a 'first-person account" is not the content of that account, but the fact that one has learned to use it as the vehicle of spontaneous conceptualization in introspection and self-description.* We all of us, as children learned to use the framework of current folk psychology in this role. But it is entirely possible for a person or culture to learn and use some other framework in that role, the framework of cognitive neuroscience, perhaps. Given a deep and practiced familiarity with the developing idioms of cognitive neurobiology, we might learn to discriminate by introspection the coding vectors in our internal axonal pathways, the activation patterns across salient neural populations, and myriad other things besides.

Should that ever happen, it would then be obvious to everyone who had made the conceptual shift that a completed cognitive neuroscience would constitute not a pinched and exclusionary picture of human consciousness, one blind to the subjective dimension of self as Jackson's argument suggests. Rather, it would be the vehicle of a grand reconstruction and expansion of our subjective consciousness, since it would provide us with a conceptual framework that; unlike folk psychology, is at last equal to the kinematical and dynamical intricacies of the world within. (See also chapter 1 of this volume l and Churchland 1979, section 16.)

Real precedents for such a reformation can be drawn from our own history .We did not lose contact with a metaphysically distinct dimension of reality when we stopped seeing an immutable sparkle-strewn quintessential crystal sphere each time we looked to the heavens, and began to see instead an infinite space of gas and dust and giant stars structured by gravitational attractions and violent nuclear processes. On the contrary, we now see far more than we used to, even with the unaided eye. The diverse "colors of the stars allow us to see directly their absolute temperatures. Stellar temperature is a function of stellar mass, so we are just as reliably seeing stellar masses. The intrinsic luminosity or brightness of a star is tightly tied to these same features, and thus is also visually available, no matter how bright or faint the star may appear from Earth. Apparent brightness is visually obvious also, of course, and the contrast between the apparent and the intrinsic brightnesses gives you the star's rough distance from Earth. In this way is the character and three-dimensional distribution of complex stellar objects in a volume of interstellar space hundreds of light years on aside made visually available to your unaided eyes from your own back yard, given only the right conceptual framework for grasping it, and observational practice in using that framework. From within the new framework, one finds asystematic significance in experiential details that hitherto went largely or entirely unnoticed (compare Feyerabend 1963b).

The case of inner space is potentially the same. We will not lose contact with a metaphysically distinct dimension of self when we stop introspecting inarticulable qualia, and start introspecting "instead" sensory coding vectors and sundry activation patterns within the vector spaces of our accessible cortical areas. As with the revolution in astronomy, the prospect is one we should welcome as metaphysically liberating, rather than deride as metaphysically irrelevant or metaphysically impossible.