Matter and Consciousness

Paul Churchland, 1984

Chapter 2: The Ontological Problem (the Mind-Body Problem)

3. Reductive Materialism (The Identity Theory))

Reductive materialism, more commonly known as *the identity theory*, is the most straightforward of the several materialist theories of mind. Its central claim is simplicity itself: Mental states *are* physical states of the brain. That is, each type of mental state or process is *numerically identical with* (is one and the very same thing as) some type of physical state or process within the brain or central nervous system. At present we do not know enough about the intricate functionings of the brain actually to state the relevant identities, but the identity theory is committed to the idea that brain research will eventually reveal them ....

Historical Parallels

As the identity theorist sees it, the result here predicted has familiar parallels elsewhere in our scientific history. Consider sound. We now know that sound is just a train of compression waves traveling through the air, and that the property of being high pitched is identical with the property of having a high oscillatory frequency. We have learned that light is just electromagnetic waves, and our best current theory says that the color of an object is identical with a triplet of reflectance efficiencies the object has, rather like a musical chord that it strikes, though the 'notes' are struck in electromagnetic waves instead of in sound waves. We now appreciate that the warmth or coolness of a body is just the energy of motion of the molecules that make it up: warmth is identical with high average molecular kinetic energy, and coolness is identical with low average molecular kinetic energy. We know that lightning is

identical with a sudden large-scale discharge of electrons between clouds, or between the atmosphere and the ground. What we now think of as 'mental states,' argues the identity theorist, are identical with brain states in exactly the same way.

Intertheoretic Reduction

These illustrative parallels are all cases of successful *intertheoretic reduction.* That is, they are all cases where a new and very powerful theory turns out to entail a set of propositions and principles that mirror perfectly (or almost perfectly) the propositions and principles of some older theory or conceptual framework. The relevant principles entailed by the new theory have the same structure as the corresponding principles of the old framework, and they apply inexactly the same cases. The only difference is that where the old principles contained (for example) the notions of "heat," "is hot," and "is cold," the new principles contain instead the notions of "total molecular kinetic energy," "has a high mean molecular kinetic energy," and "has a low mean molecular kinetic energy."

If the new framework is far better than the old at explaining and predicting phenomena, then we have excellent reason for believing that the theoretical terms of the *new* framework are the terms that describe reality correctly. But if the old framework worked adequately, so far as it went, and if it parallels a portion of the new theory in the systematic way described, then we may properly conclude that the old terms and the new terms refer to the very same things, or express the very same properties. We conclude that we have apprehended the very same reality that is incompletely described by the old framework, but with a new and more penetrating conceptual framework. And we announce what philosophers of science call "intertheoretic identities": light *is* electromagnetic waves, temperature *is* mean molecular kinetic energy, and so forth.

The examples of the preceding two paragraphs share one more important feature in common. They are all cases where the things or properties on the receiving end of the reduction are *observable* things and properties within our *common-sense* conceptual framework. They show that

intertheoretic reduction occurs not only between conceptual frameworks in the theoretical stratosphere: common-sense observables can also be reduced. There would therefore be nothing particularly surprising about a reduction of our familiar introspectible mental states to physical states of the brain. All that would be required would be that an explanatorily successful neuroscience develop to the point where it entails a suitable "mirror image" of the assumptions and principles that constitute our common-sense conceptual framework for mental states, an image where brain-state terms occupy the positions held by mental-state terms in the assumptions and principles of common sense. If this (rather demanding) condition were indeed met, then, as in the historical cases cited, we would be justified in announcing a reduction, and in asserting the identity of mental states with brain states.

Arguments for the Identity Theory

What reasons does the identity theorist have for believing that neuroscience will eventually achieve the strong conditions necessary for the reduction of our "folk" psychology? There are at least four reasons, all directed at the conclusion that the correct account of human-behavior-and-its-causes must reside in the physical neurosciences.

We can point first to the purely physical origins and ostensibly physical constitution of each individual human. One begins as a genetically programmed monocellular organization of molecules (a fertilized ovum), and one develops from there by the accretion of further molecules whose structure and integration is controlled by the information coded in the DNA molecules of the cell nucleus. The result of such a process would be a purely physical system whose behavior arises from its internal operations and its interactions with the rest of the physical world. And those behavior-controlling internal operations are precisely what the neurosciences are about.

This argument coheres with a second argument. The origins of each *type* of animal also appear exhaustively physical in nature. The argument from evolutionary history discussed earlier ... lends further support to the identity theorist's claim, since evolutionary theory provides the only

serious explanation we have for the behavior-controlling capacities of the brain and central nervous system. Those systems were selected for because of the many advantages (ultimately, the reproductive advantage) held by creatures whose behavior was thus controlled. Again our behavior appears to have its basic causes in neural activity.

The identity theorist finds further support in the argument, discussed earlier, from the neural dependence of all known mental phenomena ... . This is precisely what one should expect, if the identity theory is true. Of course, systematic neural dependence is also a consequence of property dualism, but here the identity theorist will appeal to considerations of simplicity. Why admit two radically different classes of properties and operations if the explanatory job can be done by one?

A final argument derives from the growing success of the neurosciences in unraveling the nervous systems of many creatures and in explaining their behavioral capacities and deficits in terms of the structures discovered. The preceding arguments all suggest that neuroscience should be successful in this endeavor, and the fact is that the continuing history of neuroscience bears them out. Especially in the case of very simple creatures (as one would expect), progress has been rapid. And progress has also been made with humans, though for obvious moral reasons exploration must be more cautious and circumspect. In sum, the neurosciences have a long way to go, but progress to date provides substantial encouragement to the identity theorist.

Even so, these arguments are far from decisive in favor of the identity theory. No doubt they do provide an overwhelming case for the idea that the causes of human and animal behavior are essentially physical in nature, but the identity theory claims more than just this. It claims that neuroscience will discover a taxonomy of neural states that stand in a one-to-one correspondence with the mental states of our common-sense taxonomy. Claims for intertheoretic identity will be justified only if such a match-up can be found. But nothing in the preceding arguments guarantees that the old and new frameworks will match up in this way, even if the new framework is a roaring success at explaining and predicting our behavior. Furthermore, there are arguments from other positions within the materialist camp to the effect that such convenient

match-ups are rather unlikely. Before exploring those, however, let us look at some more traditional objections to the identity theory.

Arguments Against the Identity Theory

We may begin with the argument from introspection discussed earlier. Introspection reveals a domain of thoughts, sensations, and emotions, not a domain of electrochemical impulses in a neural network. Mental states and properties, as revealed in introspection, appear radically different from any neurophysiological states and properties. How could they possibly be the very same things? The answer, as we have already seen, is, "Easily." In discriminating red from blue, sweet from sour, and hot from cold, our external senses are actually discriminating between subtle differences in intricate electromagnetic, stereochemical, and micromechanical properties of physical objects. But our senses are not sufficiently penetrating to reveal on their own the detailed nature of those intricate properties. That requires theoretical research and experimental exploration with specially designed instruments. The same is presumably true of our "inner" sense: introspection. It may discriminate efficiently between a great variety of neural states, without being able to reveal on its own the detailed nature of the states being discriminated. Indeed, it would be faintly miraculous if it did reveal them, just as miraculous as if unaided sight were to reveal the existence of interacting electric and magnetic fields whizzing by with an oscillatory frequency of a million billion hertz and a wavelength of less than a millionth of a meter. For despite 'appearances,' that is what light is. The argument from introspection, therefore, is quite without force.

The next objection argues that the identification of mental states with brain states would commit us to statements that are literally unintelligible, to what philosophers have called "category errors," and that the identification is therefore a case of sheer conceptual confusion. We may begin the discussion by noting a most important law concerning numerical identity.[1] Leibniz' Law states that two items are numerically

[1] [RJ: *Numerical* identity is a trivial 'relation', since it only relates each object to *itself*. Thus Henrik and Daniel Sedin are not *numerically* identical, since they

identical just in case any property had by either one of them is also had by the other: in logical notation,

$$\forall x \, \forall y(\, x = y \; \leftrightarrow \; \forall F \, (F(x) \leftrightarrow F(y)) \,)$$

This law suggests a way of refuting the identity theory: find some property that is true of brain states, but not of mental states (or vice versa), and the theory would be exploded.

Spatial properties were often cited to this end. Brain states and processes must of course have some specific spatial location: in the brain as a whole, or in some part of it. And if mental states are identical with brain states, then they must have the very same spatial location. But it is literally meaningless, runs the argument, to say that my feeling-of-pain is located in my ventral thalamus, or that my belief-that-the-sun is-a-star is located in the temporal lobe of my left cerebral hemisphere. Such claims are as meaningless as the claim that the number 5 is green, or that love weighs twenty grams.

Trying the same move from the other direction, some have argued that it is senseless to ascribe the various *semantic* properties to brain states. Our thoughts and beliefs, for example, have a meaning, a specific propositional content; they are either true or false; and they can enjoy relations such as consistency and entailment. If thoughts and beliefs were brain states, then all these semantic properties would have to be true of brain states. But it is senseless, runs the argument, to say that some resonance in my association cortex is true, or logically entails some other resonance close by, or has the meaning that *P*.

Neither of these moves has the same bite it did twenty years ago, since familiarity with the identity theory and growing awareness of the brain's role have tended to reduce the feelings of semantic oddity produced by the claims at issue. But even if they still struck all of us as semantically confused, this would carry little weight. The claim that sound has a wavelength, or that light has a frequency, must have seemed equally

are two different people. (If Henrik passes to Daniel, then he is not passing to himself.) On the other hand, Superman (in the story) is numerically identical to Clark Kent.]

unintelligible in advance of the conviction that both sound and light are wave phenomena. (See, for example, Bishop Berkeley's eighteenth-century dismissal of the idea that sound is a vibratory motion of the air, in Dialogue I of his *Three Dialogues.* The objections are voiced by Philonous.) The claim that warmth is measured in kilogram meters^2 /seconds^2 would have seemed semantically perverse before we understood that temperature is mean molecular kinetic energy. And Copernicus' sixteenth-century claim that the earth *moves* also struck people as absurd to the point of perversity. It is not difficult to appreciate why. Consider the following argument.

> Copernicus' claim that the earth moves is sheer conceptual confusion. For consider what it *means* to say that something moves: "*x* moves" means "*x* changes position relative to the earth." Thus, to say that the earth moves is to say that the earth changes position relative to itself! Which is absurd. Copernicus' position is therefore an abuse of language.

The *meaning analysis* here invoked might well have been correct, but all that would have meant is that the speaker should have set about changing his meanings. The fact is, any language involves a rich network of assumptions about the structure of the world, and if a sentence *S* provokes intuitions of semantic oddness, that is usually because S violates one or more of those background assumptions. But one cannot always reject *S* for that reason alone, since the overthrow of those background assumptions may be precisely what the facts require. The "abuse" of accepted modes of speech is often an essential feature of real scientific progress! Perhaps we shall just have to get used to the idea that mental states have anatomical locations and brain states have semantic properties.

While the charge of sheer senselessness can be put aside, the identity theorist does owe us some account of exactly how physical brain states can have semantic properties. The account currently being explored can be outlined as follows. Let us begin by asking how it is that a particular *sentence* (= utterance type) has the specific propositional content it has: the sentence "La pomme est rouge," for example. Note first that a sentence is always an integrated part of an entire system of sentences: a language. Any given sentence enjoys many relations with countless other

sentences: it entails many sentences, is entailed by many others, is consistent with some, is inconsistent with others, provides confirming evidence for yet others, and so forth. And speakers who use that sentence within that language draw inferences in accordance with those relations. Evidently, each sentence (or each set of equivalent sentences) enjoys a unique pattern of such entailment relations: it plays a distinct inferential role in a complex linguistic economy. Accordingly, we say that the sentence "La pomme est rouge" has the propositional content, *the apple is red,* because the sentence "La pomme est rouge" plays *the same role* in French that the sentence "The apple is red" plays in English. To have the relevant propositional content is just to play the relevant inferential role in a cognitive economy.

Returning now to types of brain states, there is no problem in principle in assuming that one's brain is the seat of a complex inferential economy in which types of brain states are the role-playing elements. According to the theory of meaning just sketched, such states would then have propositional content, since having content is not a matter of whether the contentful item is a pattern of sound, a pattern of letters on paper, a set of raised Braille bumps, or a pattern of neural activity. What matters is the inferential role the item plays. Propositional content, therefore, seems within the reach of brain states after all.

We began this subsection with an argument against materialism that appealed to the qualitative *nature* of our mental states, as revealed in introspection. The next argument appeals to the simple fact that they are introspectible at all.

1. My mental states are introspectively known by me as states of my conscious self.

2. My brain states are *not* introspectively known by me as states of my conscious self.

Therefore, by Leibniz' Law (that numerically identical things must have exactly the same properties),

3. My mental states are not identical with my brain states.

This, in my experience, is the most beguiling form of the argument from introspection, seductive of freshmen and faculty alike. But it is a straightforward instance of a well-known fallacy, which is clearly illustrated in the following parallel arguments:

1. Muhammad Ali is widely known as a heavyweight champion.
2. Cassius Clay is *not* widely known as a heavyweight champion.

Therefore, by Leibniz' Law,

3. Muhammad Ali is not identical with Cassius Clay.

or,

1. Aspirin is recognized by John to be a pain reliever.
2. Acetylsalicylic acid is *not* recognized by John to be a pain reliever.

Therefore, by Leibniz' Law,

3. Aspirin is not identical with acetylsalicylic acid.

Despite the truth of the relevant premises, both conclusions are false: the identities are wholly genuine. Which means that both arguments are invalid. The problem is that the "property" ascribed in premise (1), and withheld in premise (2), consists only in the subject item's being *recognized, perceived,* or *known* as something-or-other. But such apprehension is not a genuine property of the item itself, fit for divining identities, since one and the same subject may be successfully recognized under one name or description, and yet fail to be recognized under another (accurate, coreferential) description. Bluntly, Leibniz' Law is not valid for these bogus "properties." The attempt to use them as above commits what logicians call an *intensional* fallacy. The premises may reflect, not the failure of certain objective identities, but only our continuing failure to appreciate them.

A different version of the preceding argument must also be considered, since it may be urged that one's brain states are more than merely not (yet) known by introspection: they are not *knowable* by introspection under any circumstances. Thus,

1. My mental states are knowable by introspection.
2. My brain states are *not* knowable by introspection.

Therefore, by Leibniz' Law,

3. My mental states are not identical with my brain states.

Here the critic will insist that being know*able* by introspection *is* a genuine property of a thing, and that this modified version of the argument is free of the "intensional fallacy" discussed above.

And so it is. But now the materialist is in a position to insist that the argument contains a false premise—premise (2). For if mental states are indeed brain states, then it is really brain states we have been introspecting all along, though without fully appreciating what they are. And if we can learn to think of and recognize those states under mentalistic descriptions, as we all have, then we can certainly learn to think of and recognize them under their more penetrating neurophysiological descriptions. At the very least, premise (2) simply begs the question against the identity theorist. The mistake is amply illustrated in the following parallel argument:

1. Temperature is knowable by feeling.
2. Mean molecular kinetic energy is *not* knowable by feeling.

Therefore, by Leibniz' Law,

3. Temperature is not identical with mean molecular kinetic energy.

This identity, at least, is long established, and this argument is certainly unsound: premise (2) is false. 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 what our discriminatory mechanisms are keyed to. Perhaps our brain states are similarly accessible ....

Consider now a final argument, again based on the introspectible qualities of our sensations. Imagine a future neuroscientist who 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. If for some reason she has never actually *had* a sensation-of-red (because of color blindness, say, or an unusual environment), then there will remain something she does *not* know about certain sensations: *what it* is *like to have a sensation-of-red.* Therefore, complete knowledge of the physical facts of visual perception and its related brain activity still leaves something out. Accordingly, materialism cannot give an adequate account of all mental phenomena, and the identity theory must be false.

The identity theorist can reply that this argument exploits an unwitting equivocation on the term "know." Concerning our neuroscientist's utopian knowledge of the brain, "knows" means something like "has mastered the relevant set of neuroscientific propositions." Concerning her (missing) knowledge of what it is like to have a sensation-of-red, "knows" means something like "has a prelinguistic representation of redness in her mechanisms for noninferential discrimination." It is true that one might have the former without the latter, but the materialist is not committed to the idea that having knowledge in the former sense automatically constitutes having knowledge in the second sense. The identity theorist can admit a duality, or even a plurality, of different *types of knowledge* without thereby committing himself to a duality in *types of things known.* The difference between a person who knows all about the visual cortex but has never enjoyed the 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, nonphysical *qualia* by the latter), but rather in the different *type,* or *medium,* or *level* of representation each has of exactly the same thing: brain states.

In sum, there are pretty clearly more was of "having knowledge" than just having mastered a set of sentences, and the materialist can freely admit that one has "knowledge" of one's sensations in a way that is independent of the neuroscience one may have learned. Animals, including humans, presumably have a prelinguistic mode of sensory representation. 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 mere storage of sentences.* All the

identity theorist needs to claim is that those other modes of representation will also yield to neuroscientific explanation.

The identity theory has proved to be very resilient in the face of these predominantly antimaterialist objections. But further objections, rooted in competing forms of materialism, constitute a much more serious threat, as the following sections will show.