This essay is chapter 18 of Concealment and Exposure and Other Essays (*New York, Oxford University Press, 2002). An earlier version appeared in* New Essays on the A Priori*, Paul Boghossian and Christopher Peacocke, eds. (Oxford, Clarendon Press, 2000).*

THE PSYCHOPHYSICAL NEXUS

Thomas Nagel

I. The Mind-Body Problem after Kripke

This essay will explore an approach to the mind-body problem that is distinct both from dualism and from the sort of conceptual reduction of the mental to the physical that proceeds via causal behaviorist or functionalist analysis of mental concepts. The essential element of the approach is that it takes the subjective phenomenological features of conscious experience to be perfectly real and not reducible to anything else--but nevertheless holds that their systematic relations to neurophysiology are not contingent but necessary.

A great deal of effort and ingenuity has been put into the reductionist program, and there have been serious attempts in recent years to accommodate within a functionalist framework consciousness and phenomenological qualia in particular.[1] The

[1] See for example Sydney Shoemaker, "Self-Knowledge and 'inner sense,' Lecture III: The phenomenal character of experience," in The First-person Perspective and Other Essays (Cambridge University Press, 1996). I will use the term "functionalism" throughout this essay in

effort has produced results that reveal a good deal that is true about the relations between consciousness and behavior, but not an account of what consciousness is. The reason for this failure is unsurprising and always the same. However complete an account may be of the functional role of the perception of the color red in the explanation of behavior, for example, such an account taken by itself will have nothing to say about the specific subjective quality of the visual experience, without which it would not be a conscious experience at all.

If the intrinsic character of conscious experience remains stubbornly beyond the reach of contextual, relational, functional accounts, an alternative strategy seems called for. The exploration of such an alternative should be of interest even to those who remain convinced that functionalism is the right path to follow, since philosophical positions can be evaluated only by comparison with the competition. The alternative I wish to explore can be thought of as a response to the challenge issued by Saul Kripke at the end of Naming and Necessity:

> That the usual moves and analogies are not available to solve the problems of the identity theorist is, of course, no proof that no moves are available....I suspect, however, that the present considerations tell heavily against the usual forms of materialism. Materialism, I think, must hold that a physical description of the

---

an unsophisticated way, to refer to theories that identify mental states by their typical causal roles in the production of behavior – also called their "functional" roles. I shall leave aside the version of functionalism that identifies mental states with computational states.

world is a complete description of it, that any mental facts are 'ontologically dependent' on physical facts in the straightforward sense of following from them by necessity. No identity theorist seems to me to have made a convincing argument against the intuitive view that this is not the case.[2]

Kripke's view of functionalism and causal behaviorism is the same as mine: that the inadequacy of these analyses of the mental is self-evident. He does not absolutely rule out a form of materialism that is not based on such reductionist analyses, but he says that it has to defend the very strong claim that mental phenomena are strictly necessary consequences of the operation of the brain--and that the defense of this claim lies under the heavy burden of overcoming the prima facie modal argument that consciousness and brain states are only contingently related, since it seems perfectly conceivable about any brain state that it should exist exactly as it is, physically, without any accompanying consciousness. The intuitive credibility of this argument, which descends from Descartes' argument for dualism, is considerable. It appears at first blush that we have a clear and distinct enough grasp on both phenomenological consciousness and physical brain processes to see that there can be no necessary connection between them.

That is the position that I hope to challenge. It seems to me that post-Kripke, the most promising line of attack on the mind-body problem is to see whether any sense can be made of the idea that mental processes might be physical processes necessarily but not analytically. I would not, however, try to defend the claim that "a physical description of the world is a complete description of it," so my position is not a form of materialism in

[2] Saul Kripke, Naming and Necessity (Harvard University Press, 1980), p. 155.

Kripke's sense. It is certainly not a form of physicalism. But there may be other forms of noncontingent psychophysical identity. So I shall argue.

Because I am going to be talking about different kinds of necessity and contingency throughout the argument, I should say something at the outset about my assumptions, which will not be universally shared. The set of ideas about necessity and contingency with which I shall be working derives largely from Kripke. This means that the semantic category of analytic or conceptual truths, the epistemological category of a priori truths, and the metaphysical category of necessary truths do not coincide--nor do their complements: synthetic, a posteriori, and contingent truths.

I believe that there are conceptual truths, and that they are discoverable a priori, through reflection by a possessor of the relevant concepts--usually with the help of thought experiments--on the conditions of their application. Often the process of discovery will be difficult, and the results controversial. Conceptual truths may or may not be necessary truths. In particular, conceptual truths about how the reference of a term is fixed may identify contingent properties of the referent, though these are knowable a priori to a possessor of the concept.

Not everything discoverable a priori is a conceptual truth--for example the calculation of the logical or mathematical consequences that follow from a set of theoretical premises is a priori, but not, I would say, conceptual. And while some conceptual truths are necessary, not all necessary truths are conceptual. This applies not only to mathematical or theoretical propositions discoverable by a priori reasoning, but also, as Kripke showed, to certain identity statements that cannot be known a priori, such

as the identity of heat with molecular motion or that of water with H2O.

The relations among these different types of truths are intricate. In the case of the identity of water with H2O, for example, as I shall explain more fully later, the following appears to hold. First, there are some conceptual truths about water--its usual manifest physical properties under the conditions that prevail in our world. These are the properties by which we fix the reference of the term "water," and they are knowable a priori. Most of them are contingent properties of water, because they depend on other things as well, but some of them may be necessary, if they follow from the intrinsic nature of water alone. Second, there are theoretical truths, derivable from principles of chemistry and physics, about the macroscopic properties, under those same conditions, of the compound H2O. These are necessary consequences of premises which are partly necessary (the nature of hydrogen and oxygen) and partly contingent. Third, there is the a posteriori conclusion, from evidence that the manifest properties of the water with which we are acquainted are best explained in this way, that water is in fact nothing but H2O. This is a necessary truth, though discovered a posteriori, because if it is true then any other substance with the same manifest properties which did not consist of H2O would not be water. And this last conditional clause, following "because," is a conceptual truth, discoverable by reflection on what we would say if we encountered such a substance.

In the context with which we are concerned here, the mind-body problem, functionalism is the claim that it is a conceptual truth that any creature is conscious, and is the subject of various mental states, if and only if it satisfies certain purely structural conditions of the causal organization of its behavior and interaction with the

environment--whatever may be the material in which that organization is physically (or nonphysically) realized. I do not believe that this is a conceptual truth, because I do not believe that the conceptual implication from functional organization to consciousness holds.

I don't doubt that all the appropriately behaved and functionally organized creatures around us are conscious, but that is something we know on the basis of evidence, not on the basis of conceptual analysis. It may even be impossible in fact for a creature to function in these ways without consciousness; but if so, it is not a conceptual impossibility but some other kind. The functional organization of purely physical behavior, without more, is not enough to entail that the organism or system has subjective conscious experience, with experiential qualities. I make this claim particularly about sensations and the other qualities of sentience, rather than about higher-order intentional states like belief or desire--though I am inclined to think that they too require at least the capacity for sentience. My rejection of functionalism is based on the conviction that the subjective qualitative character of experience--what it is like for its subject--is not included or entailed by any amount of behavioral organization, and that it is a conceptually necessary condition of conscious states that they have some such character.

On the other hand, I will argue later that there <u>is</u> a conceptual connection between consciousness and behavioral or functional organization, but in the opposite direction. I deny the functionalist biconditional because of the falsity of one of its conjuncts, but I think a weak version of the opposite conjunct is true. I believe it is a conceptual truth about the visual experience of colors, for example, that it enables a physically intact

human being to discriminate colored objects by sight, and that this will usually show up in his behavior in the appropriate circumstances, provided that he meets other psychological and physical conditions. This is a conceptual truth about color vision analogous to the conceptual truths about the manifest properties of water in our world: In both cases the manifestations are contingent properties of the thing itself, dependent on surrounding circumstances. Functional organization is not a conceptually sufficient condition for mental states, but it is part of our concept of mental states that they in fact occupy something like the roles in relation to behavior that functionalists have insisted upon. Such roles permit us to fix the reference of mental terms. But they are, at least in general, contingent rather than necessary properties of the conscious mental states that occupy them.

Finally, and this is the main point, while it is obviously not conceptually necessary that conscious mental states are tied to specific neurophysiological states, I contend that there are such connections and that they hold necessarily. They are not conceptual, and they are not discoverable a priori, but they are not contingent. They belong, in other words, to the category of a posteriori necessary truths. To explain how, and to characterize the type of necessity that could hold in such a case, is the problem.

Kripke argued that if the psychophysical identity theory is to be a hypothesis analogous to other empirical reductions or theoretical identifications in science, like the identification of heat with molecular motion or fire with oxidation, it cannot be a contingent proposition. It must be necessarily true if true at all, since a theoretical identity statement tells us what something <u>is</u>, not just what happens to be true of it. In the

vocabulary introduced by Kripke, the terms of such an identity are both rigid designators, and they apply or fail to apply to the same things in all possible worlds.

Kripke observes that there is an appearance of contingency even in the standard cases of theoretical identity. The identification of heat with molecular motion is not analytic, and it cannot be known a priori. It may seem that we can easily conceive of a situation in which there is heat without molecular motion, or molecular motion without heat. But Kripke points out that this is a subtle mistake. When one thinks one is imagining heat without molecular motion, one is really imagining the feeling of heat being produced by something other than molecular motion. But that would not be heat--it would merely be a situation epistemically indistinguishable from the perception of heat. "Heat," being a rigid designator, refers to the actual physical phenomenon that is in fact responsible for all the manifestations on the basis of which we apply the concept in the world as it is. The term refers to that physical phenomenon and to no other, even in imagined situations where something else is responsible for similar appearances and sensations. This is so because the appearances and sensations of heat are not themselves heat, and can be imagined to exist without it.

Kripke then points out that a similar strategy will not work to dissipate the appearance of contingency in the case of the relation between sensations and brain processes. If I seem to be able to imagine the taste of chocolate in the absence of its associated brain process, or the brain process unaccompanied by any such experience, we cannot say that this is merely to imagine the <u>appearance</u> of the experience without the experience, or vice versa. There is, in this case, no way of separating the thing itself from

the way it appears to us, as there is in the case of heat. We identify experiences not by their contingent effects on us, but by their intrinsic phenomenological qualities. So if they are really identical with physical processes in the brain, the vivid appearance that we can clearly conceive of the qualities without the brain processes, and vice versa, must be shown to be erroneous in some other way.

My hope is to show that this can be done, without abandoning a commitment to the reality of the phenomenological content of conscious experience. If the appearance of contingency in the mind-body relation can be shown to be illusory, or if it can be shown how it might be illusory, then the modal argument against some sort of identification will no longer present an immovable obstacle to the empirical hypothesis that mental processes are brain processes.

The hypothesis would resemble familiar theoretical identities, like that between heat and molecular motion, in some respects but not in others. It would be nonanalytic, discoverable only a posteriori, and necessarily true if true. But of course it could not be established by discovering the underlying physical cause of the appearance of conscious experience, on analogy with the underlying physical cause of the appearance of heat--since in the case of experience, the appearance is the thing itself and not merely its effect on us.

Clearly this would require something radical. We cannot at present see how the relation between consciousness and brain processes might be necessary. The logical gap between subjective consciousness and neurophysiology seems unbridgeable, however close may be the contingent correlations between them. To see the importance of this

gap, consider how the necessary connection is established in other cases.

To show that water is H2O or that heat is molecular motion, it is necessary to show that the chemical or physical equivalence can account fully and exhaustively for everything that is included in the ordinary prescientific concepts of water and heat--the manifest properties on the basis of which we apply those concepts. Not only must the scientific account explain causally all the external effects of water or heat, such as their effects on our senses. It must also account in a more intimate manner for their familiar intrinsic properties, revealing the true basis of those properties by showing that they are entailed by the scientific description. Thus, the density of water, its passage from solid to liquid to gas at certain temperatures, its capacity to enter into chemical reactions or to appear as a chemical product, its transparency, viscosity, electrical conductivity, and so forth, must all be accounted for in a particularly strong way by its chemical analysis as H2O, together with whatever laws govern the behavior of such a compound. In brief, the essential intrinsic properties of water on the macro level must be properties that simply follow from the behavior of H2O under normal conditions. Otherwise it will not be possible to say that water is constituted of H2O and nothing else.

In what sense must the familiar, manifest properties of water follow from the properties of H2O to support the claim of constitution? To require a strict logical entailment would be far too demanding. We do not find that even in the case of reduction of one scientific theory to another, more fundamental theory. There is always a certain amount of slippage and deviation around the edges. But what we can expect is that the reducing theory will entail something close enough to the familiar properties of the thing to be reduced, allowing for the roughness of ordinary concepts and perceptual observations, to permit us to conclude that nothing more is

needed to explain why H2O, for example, has the macroscopic features of water.

To illustrate: One reason for the absence of strict entailment is that the relation between the physics of H2O and the macroscopic properties of water is probabilistic. It is, I am assured by those who know more about these matters than I, physically possible for H2O to be a solid at room temperature, though extremely unlikely. That means that if water is H2O, it is possible for water to be a solid at room temperature. And similar things can apparently be said about the other manifest properties of water by means of which the reference of the term is fixed. Yet I think these esoteric facts do not remove the element of necessity in the relation between the properties of H2O and the macroscopic properties conceptually implied by our concept of water. Those macroscopic, manifest properties are not really inconsistent with an interpretation under which they are merely probabilistic, provided the probabilities are so astronomically high that their failure is for all practical purposes impossible, and it would never be rational to believe that it had occurred. It is enough if the physics of H2O entails that the probability of water having these properties under normal background conditions is so close to 1 as makes no experiential difference. Let me take this qualification as understood when I speak of entailment from now on.3

---

3 One further point: Even if there are laws governing the behavior of molecules in large numbers that are genuinely higher-order and not merely the statistical consequences of the probabilistic or deterministic laws governing the individual particles – holistic laws, so to speak – it still does not affect the point. Facts about the macroscopic properties of a substance like water, or an event like a thunderstorm, would still be constitutively entailed by the facts about the behavior of the microscopic or submicroscopic constituents – whatever kinds of laws might

This rough variety of "upward entailment" is a necessary condition of any successful scientific reduction in regard to the physical world. It is the a priori element in a posteriori necessary theoretical identities. We begin with an ordinary concept of a natural kind or natural phenomenon. This concept--heat or water--refers to the actual examples to which we apply it, and with which we are in some kind of direct or indirect contact through our occupation of the world. To establish that those examples are in fact identical with something not directly manifest to perception but describable only by atomic theory, we must show that the prescientifically familiar intrinsic features of heat and water are nothing but the gross manifestations of the properties of these physicochemical constituents--that the liquidity of water, for example, consists simply of a certain type of movement of its molecules with respect to one another. If the properties of the substance that we refer to by the term "water" can be exhaustively accounted for by such a micro-analysis, and if experiment confirms that this is in fact the situation that obtains, then that tells us what water really is.

The result is a posteriori because it requires not only the a priori demonstration that H2O could account for the phenomena, but empirical confirmation that this and not something else is what actually underlies the manifest properties of the substance we refer to as water. That would come from experimental confirmation of previously unobserved implications of the hypothesis, and disconfirmation of the implications of alternative hypotheses, e.g. that water is an element. Thus it is not a conceptual reduction.

---

be required to account for this behavior.

Nevertheless it is a necessary identity because our concept of water refers to the actual water around us, whatever it is, and not to just any substance superficially resembling water. If there could be something with the familiar manifest properties of water which was not H2O, it would not be water. But to reach this conclusion, we must see that the behavior of H2O provides a true and complete account, with nothing left out—an approximate entailment--of the features that are conceptually essential to water, and that this account is in fact true of the water around us.

It is this "upward entailment" that is so difficult to imagine in the case of the corresponding psychophysical hypothesis, and that is the nub of the mind-body problem. We understand the entailment of the liquidity of water by the behavior of molecules through geometry, or more simply the micro-macro or part-whole relation. Something analogous is true of every physical reduction, even though the spatiotemporal framework can be very complicated and hard to grasp intuitively. But nothing like this will help us with the mind-body case, because we are not dealing here merely with larger and smaller grids. We are dealing with a gap of a totally different kind, between the objective spatiotemporal order of the physical world and the subjective phenomenological order of experience. And here it seems clear in advance that no amount of physical information about the spatiotemporal order will entail anything of a subjective, phenomenological character. However much our purely physical concepts may change in the course of further theoretical development, they will all have been introduced to explain features of the objective spatiotemporal order, and will not have implications of this radically different logical type.

But without an upward entailment of some kind, we will not have a proper reduction, because in any proposed reduction of the mental to the physical, something will have been left out--something essential to the phenomenon being reduced. Unless this obstacle can be overcome, it will be impossible to claim that the relation between sensations and brain processes is analogous to the relation between heat and molecular motion--a necessary but a posteriori identity.

Yet I believe that is the region in which the truth probably lies. The evident massive and detailed dependence of what happens in the mind on what happens in the brain provides, in my view, strong evidence that the relation is not contingent but necessary. It cannot take the form of a reduction of the mental to the physical, but it may be necessary all the same. The task is to try to understand how that might be the case.[4]

II. Subjectivity and the Conceptual Irreducibility of Consciousness

The source of the problem--what seems to put such a solution out of reach--is the lack of any intelligible internal relation between consciousness and its physiological basis. The apparent conceivability of what in current philosophical jargon is known as a "zombie"--i.e. an exact physiological and behavioral replica of a living human being that nevertheless has no consciousness--may not show that such a thing is possible, but it does

[4] My position is very like that of Colin McGinn, but without his pessimism. See The Problem of Consciousness (Blackwell, 1991). What I have to say here is also a development of a suggestion in The View From Nowhere (Oxford University Press, 1986), pp. 51-53.

show something about our concepts of mind and body. It shows that those concepts in their present form are not logically connected in such a way that the content of the idea of consciousness is exhausted by a physical or behavioral-functional specification.

But the rejection of conceptual reduction is only the beginning of the story. The problem is to look for an alternative account of the evidently very close relation between consciousness and the brain which does not in any way accord a diminished reality to the immediate phenomenological qualities of conscious experience. Because of the causal role of mental events in the physical world, and their association with specific organic structures and processes, Cartesian dualism is implausible. Physicalism, in the sense of a complete conceptual reduction of the mental to the physical, is not a possibility, since it in effect eliminates what is distinctive and undeniable about the mental. Ostensibly weaker forms of physicalism seem always to collapse into behavioristic reductionism.

For that reason I have occasionally been drawn to some kind of property dualism; but like substance dualism, it seems just to be giving a name to a mystery, and not to explain anything: Simply to say that mental events are physical events with additional, nonphysical properties is to force disparate concepts together without thereby making the link even potentially intelligible. It suggests pure emergence, which explains nothing. But I believe these dead ends are not exhaustive, and that starting from our present concepts of mind and body, another approach is possible.

When we try to reason about the possible relations between things, we have to rely on our conceptual grasp of them. The more adequate the grasp, the more reliable our reasoning will be. Sometimes a familiar concept clearly allows for the possibility that

what it designates should also have features not implied by the concept itself--often features very different in kind from those directly implied by the concept. Thus ordinary prescientific concepts of kinds of substances, such as water or gold or blood, are in themselves silent with regard to the microscopic composition of those substances but nevertheless open to the scientific discovery, often by very indirect means, of such facts about their true nature. If a concept refers to something that takes up room in the spatiotemporal world, it provides a handle for all kinds of empirical discoveries about the inner constitution of that thing.

On the other hand, sometimes a familiar concept clearly excludes the possibility that what it designates has certain features: for example we do not need a scientific investigation to be certain that the number 379 does not have parents. There are various other things that we can come to know about the number 379 only by mathematical or empirical investigation, such as what its factors are, or whether it is greater than the population of Chugwater, Wyoming, but we know that it does not have parents just by knowing that it is a number. If someone rebuked us for being closed-minded, because we can't predict in advance what future scientific research might turn up about the biological origins of numbers, he would not be offering a serious ground for doubt.

The case of mental processes and the brain is intermediate between these two. Descartes thought it was closer to the second category, and that we could tell just by thinking about it that the human mind was not an extended material thing and that no extended material thing could be a thinking subject. But this is, to put it mildly, not nearly as self-evident as that a number cannot have parents. What does seem true is that

the concept of a mind, or of a mental event or process, fails to plainly leave space for the possibility that what it designates should turn out also to be a physical thing or event or process, as the result of closer scientific investigation--in the way that the concept of blood leaves space for discoveries about its composition. The trouble is that mental concepts don't obviously pick out things or processes that take up room in the spatiotemporal world to begin with. If they did, we could just get hold of some of those things and take them apart or look at them under a microscope. But there is a prior problem about how those concepts might refer to anything that could be subjected to such investigation: They don't give us the comfortable initial handle on the occupants of the familiar spatiotemporal world that prescientific physical substance concepts do.[5]

Nevertheless it is overconfident to conclude, from one's inability to imagine how mental phenomena might turn out to have physical properties, that the possibility can be ruled out in advance. We have to ask ourselves whether there is more behind the Cartesian intuition than mere lack of knowledge, resulting in lack of imagination.[6] Yet it is not enough merely to say, "You may be mistaking your own inability to imagine something for its inconceivability." One should be open to the possibility of withdrawing a judgment of inconceivability if offered a reason to doubt it, but there does have to <u>be</u> a

---

[5] See Colin McGinn, "Consciousness and Space," <u>Journal of Consciousness Studies</u> 2 (1995), pp. 220-30.

[6] This is the objection that Arnauld made to Descartes, in the fourth set of objections to the <u>Meditations</u>.

reason, or at least some kind of story about how the illusion of inconceivability may have arisen.

If mental events really have physical properties, we need an explanation of why they seem to offer so little purchase for the attribution of those properties. Still, the kind of incomprehensibility here is completely different from that of numbers having parents. Mental events, unlike numbers, can be roughly located in space and time, and are causally related to physical events, in both directions. The causal facts are strong evidence that mental events have physical properties, if only we could make sense of the idea.[7]

Consider another case where the prescientific concept did not obviously allow for the possibility of physical composition or structure--the case of sound. Before the discovery that sounds are waves in air or another medium, the ordinary concept permitted sounds to be roughly located, and to have properties like loudness, pitch, and duration. The concept of a sound was that of an objective phenomenon that could be heard by different people, or that could exist unheard. But it would have been very obscure what could be meant by ascribing to a sound a precise spatial shape and size, or an internal, perhaps microscopic, physical structure. Someone who proposed that sounds have physical parts, without offering any theory to explain this, would not have said anything understandable. One might say that in advance of the development of a physical theory of

[7] Compare Donald Davidson, "Mental Events," in his Essays on Actions and Events (Oxford University Press, 1980).

sound, the hypothesis that sounds have a physical microstructure would not have a clear meaning.

Nevertheless, at one remove, the possibility of such a development is evidently not excluded by the concept of sound. Sounds were known to have certain physical causes, to be blocked by certain kinds of obstacles, and to be perceptible by hearing. This was already a substantial amount of causal information, and it opened the way to the discovery of a physically describable phenomenon that could be identified with sound because it had just those causes and effects--particularly once further features of sound, like variations of loudness and pitch, could also be accounted for in terms of its precise physical character. Yet it is important that in advance, the idea that a sound has a physical microstructure would have had no clear meaning. One would not have known how to go about imagining such a thing, any more than one could have imagined a sound having weight. It would have been easy to mistake this lack of clear allowance for the possibility in the concept for a positive exclusion of the possibility by the concept.

The analogy with the case of mental phenomena should be clear. They too occupy causal roles, and it has been one of the strongest arguments for some kind of physicalism that those roles may prove upon investigation to be occupied by organic processes. Yet the problem here is much more serious, for an obvious reason: Identifying sounds with waves in the air does not require that we ascribe phenomenological qualities and subjectivity to anything physical, because those are features of the perception of sound, not of sound itself. By contrast, the identification of mental events with physical events requires the unification of these two types of properties in a single thing, and that remains

resistant to understanding. The causal argument for identification may make us believe that it is true, but it doesn't help us to understand it, and in my view, we really shouldn't believe it unless we can understand it.

The problem here, as with the other issue of purely conceptual reduction, lies in the distinctive first-person/third-person character of mental concepts, which is the grammatical manifestation of the subjectivity of mental phenomena. Though not all conscious beings possess language, our attribution of conscious states to languageless creatures implies that those states are of the kind that in the human case we pick out only through these distinctive concepts, concepts which the subject applies in his own case without observation of his body.

They are not pure first-person concepts: To try to detach their first-person application from the third person results in philosophical illusions. For example, from the purely first-person standpoint it seems intelligible that the subject of my present consciousness might have been created five minutes ago and all my memories, personality, etc. transferred from a previous subject in this same body to the newly created one, without any outwardly or inwardly perceptible sign--without any other physical or psychological change. If the pure first-person idea of 'I' defined an individual, that would make sense, but it seems reasonably clear that the real idea of 'I' has lost its moorings in this philosophical thought experiment. The point goes back to Kant, who argued that the subjective identity of the consciousness of myself at different times does not establish the objective identity of a subject or soul.[8]

[8] See Critique of Pure Reason, A 363-4: the Paralogisms of Pure Reason.

That is not to say that I understand just how the first person and the third form two logically inseparable aspects of a single concept--only that they do. This applies to all conscious mental states and events, and their properties. They are subjective, not in the sense that they are the subjects of a purely first-person vocabulary, but in the sense that they can be accurately described only by concepts in which nonobservational first-person and observational third-person attributions are systematically connected. Such states are modifications of the point of view of an individual subject.

The problem, then, is how something that is an aspect or element of an individual's subjective point of view could also be a physiologically describable event in the brain--the kind of thing which, considered under that description, involves no point of view and no distinctively immediate first-person attribution at all. I believe that as a matter of fact you can't have one without the other, and furthermore that the powerful intuition that it is conceivable that an intact and normally functioning physical human organism could be a completely unconscious zombie is an illusion--due to the limitations of our understanding. Nevertheless those limitations are real. We do not at present possess the conceptual equipment to understand how subjective and physical features could both be essential aspects of a single entity or process. Kant expresses roughly the same point in terms of his apparatus of phenomena and noumena:

> If I understand by soul a thinking being in itself, the question whether or not it is the same in kind as matter--matter not being a thing in itself, but merely a species of representations in us--is by its very terms illegitimate. For it is obvious that a thing in itself is of a different nature from the determinations which constitute

only its state.

> If on the other hand, we compare the thinking 'I' not with matter but with the intelligible that lies at the basis of the outer appearance which we call matter, we have no knowledge whatsoever of the intelligible, and therefore are in no position to say that the soul is in any inward respect different from it.[9]

What I want to propose, however, is that these conceptual limitations might be overcome--that there is not a perfect fit at every stage of our conceptual development between conceptual truths and necessary truths, and that this is the most probable interpretation of the present situation with respect to mind and brain: The dependence of mind on brain is not conceptually transparent but it is necessary nonetheless.

III. Necessary Truth and Conceptual Creativity

The greatest scientific progress occurs through conceptual change which permits empirically observed order that initially appears contingent to be understood at a deeper level as necessary, in the sense of being entailed by the true nature of the phenomena. Something like this must have happened at the birth of mathematics, but it is a pervasive aspect of physical science. This is the domain in which I think it is appropriate to speak of natural, as opposed to conceptual, necessity.

To take a simple and familiar example: It was observable to anyone before the

---

[9] Critique of Pure Reason, A 360. McGinn, too, remarks on the similarity of Kant's view to his own. See The Problem of Consciousness, pp. 81-82.

advent of modern chemistry that a fire will go out quickly if enclosed in a small airtight space. Given the prescientific concepts of air and fire, this was not a conceptual truth, and there would have been no way, on purely conceptual grounds, to discover that it was anything other than a strict but contingent correlation. However its very strictness should have suggested that it was not really contingent, but could be accounted for as a logical consequence of the true nature of fire and air, neither of which is fully revealed in the prescientific concepts.

This phenomenon is itself one of the evidentiary grounds for identifying fire with rapid oxidation, and air with a mixture of gases of which oxygen is one. Those identifications in turn reveal it to be a noncontingent truth that the enclosed fire will go out. The very process of oxidation that constitutes the fire eventually binds all the free oxygen in the airtight container, thus entailing its own termination. Once we develop the concepts of atomistic chemistry and physics that enable us to see what fire and air really are, we understand that it is not really conceivable that a fire should continue to burn in a small airtight space, even though our prescientific concepts did not make this evident.

The consequence is that conceivability arguments for the contingency of a correlation or the distinctness of differently described phenomena depend for their reliability on the adequacy of the concepts being employed. If those concepts do not adequately grasp the nature of the things to which they refer, they may yield deceptive appearances of contingency and nonidentity.

The mind-brain case seems a natural candidate for such treatment because what happens in consciousness is pretty clearly supervenient on what happens physically in the

brain. In the present state of our conceptions of consciousness and neurophysiology, this strict dependence is a brute fact and completely mysterious. But pure, unexplained supervenience is never a solution to a problem but a sign that there is something fundamental we don't know. If the physical necessitates the mental, there must be some answer to the question how it does so. An obviously systematic connection that remains unintelligible to us calls out for a theory.[10]

From the conceptual irreducibility of the mental to the physical, together with the empirical evidence of a connection between the mental and the physical so strong that it must be necessary, we can conclude that our mental concepts, or our physical concepts, or both, fail to capture something about the nature of the phenomena to which they refer, however accurate they may be as far as they go. The conceptual development that would be needed to reveal the underlying necessary connection is of a radical and scientifically unprecedented kind, because these two types of concepts as they now stand are not already open to the possibility that what they refer to should have a true nature of the other type.

Ordinary physical concepts, like that of fire, are candidly incomplete in what they reveal about the inner constitution of the manifest process or phenomenon to which they refer: They are open to the possibility that it should have a microstructural analysis of the

[10] A similar position is endorsed by Galen Strawson in Mental Reality (MIT Press, 1994) pp. 81-84, and by Allin Cottrell in "Tertium datur? Reflections on Owen Flanagan's Consciousness Reconsidered," Philosophical Psychology vol.8 (1995).

kind that it in fact proves to have. But nothing in the ordinary concepts of either consciousness or the brain leaves space for the possibility that they should have inner constitutions that would close the logical gap between them. Physical phenomena can be analyzed into their physical constituents, with the aid of scientific experimentation, and mental phenomena can perhaps be analyzed into their mental constituents, at least in some cases, but these two paths of analysis do not meet. The apparent conceivability of each of the correlated items without the other cannot be defused without something much more radical than the type of reduction that we are familiar with in the physical sciences.

That poses the general question of how we can attempt to develop conceptions that reflect the actual necessary connections and are therefore reliable tools for reasoning, and what determines whether there is hope of developing such concepts for a domain where we do not yet have them. After all, humans did not always have logical, geometrical, and arithmetical concepts, but had to develop them. Yet we cannot will a new conceptual framework into existence. It has to result from trying to think, in light of the evidence, about the subject we want to understand, and devising concepts that do better justice to it than the ones we have.

So how might we proceed in this case? While I am not going to follow them, there are precedents for this revisionist project: The idea that the physical description of the brain leaves out its mental essence and that we need to reform our concepts accordingly is not new. A version of it is found in Spinoza and it is at the heart of Bertrand Russell's neutral monism, expounded in The Analysis of Matter, An Outline of Philosophy, and other writings. He holds that physics in general describes only a causal

structure of events, leaving the intrinsic nature of its elements unspecified, and that our only knowledge of that intrinsic nature is in respect to certain physical events in our own brains, of which we are aware as percepts. He also holds that physics contains nothing incompatible with the possibility that all physical events, in brains or not, have an intrinsic nature of the same general type--though their specific qualities would presumably vary greatly. Here is what he says:

> There is no theoretical reason why a light-wave should not consist of groups of occurrences, each containing a member more or less analogous to a minute part of a visual percept. We cannot perceive a light-wave, since the interposition of an eye and brain stops it. We know, therefore, only its abstract mathematical properties. Such properties may belong to groups composed of any kind of material. To assert that the material must be very different from percepts is to assume that we know a great deal more than we do in fact know of the intrinsic character of physical events. If there is any advantage in supposing that the light-wave, the process in the eye, and the process in the optic nerve, contain events qualitatively continuous with the final visual percept, nothing that we know of the physical world can be used to disprove the supposition.[11]

---

[11] The Analysis of Matter (London: Allen and Unwin, 1927) pp. 263-4. For an excellent discussion and defense of Russell's and similar views, see Michael Lockwood, Mind, Brain and the Quantum (Blackwell, 1989), chap. 10. See also Grover Maxwell, "Rigid Designators and Mind-Brain Identity," in C. Wade Savage, ed., Minnesota Studies in the Philosophy of Science IX (University of Minnesota Press, 1978). Maxwell argues that it is physical rather than mental

Russell holds that both minds and bodies are logical constructions out of events. When I see the moon, my percept of the moon is one of an immense set of events, radiating out in all directions from the place where the moon is located, out of which the moon as physical object is a logical construction. The same percept also belongs to the psychologically connected set of events which constitute my mind, or mental life. And it also belongs to the set of events, centered in my skull but radiating out from there in all directions, out of which my brain as a physical object is a logical construction. (A physiologist's percept of my brain would also belong to this set, as well as to the sets constituting his mind and his brain.)

This means that the type of identification of a sensation with a brain process that Russell advocates amounts to the possibility of locating the sensation in a certain kind of causal structure--for example as the terminus of a sequence of events starting from the moon, and the origin of a sequence of events ending with the physiologist's observation of my brain. The import of describing it as a physical event is essentially relational. Its phenomenological quality is intrinsic in a way that its physical character is not.

This is a rich and interesting view, but it seems to me to solve the mind-body problem at excessive cost, by denying that physical properties are intrinsic. I believe that both mental and physical properties are intrinsic, and that this leaves an identity theory with the problem of how to understand the internal and necessary relation between them.

---

concepts that are topic-neutral, and that there is nothing to prevent their referring nonrigidly to what mental concepts designate rigidly.

The theory also leaves untouched the problem of relating the subjectivity of the mental to its physical character. Russell did have something to say about this--identifying subjectivity with dependence on the specific character of the individual's brain--but I don't think it is sufficient.

Russell's view that the intrinsic nature of physical brain processes is mental would certainly explain why the apparent conceivability of a zombie was an illusion, but it seems to me not to account for the necessity of the mind-body relation in the right way. I am sympathetic to the project of reducing both the physical and the mental to a common element, but this is too much like reducing the physical to the mental.

More recent forms of reductionism are unsatisfactory in other ways. Even if we interpret the physicalist-functionalist movement in philosophy of mind as a form of conceptual revisionism rather than analysis of what our ordinary concepts already contain, I believe it has failed because it is too conservative: It has tried to reinterpret mental concepts so as to make them tractable parts of the framework of physical science. What is needed is a search for something more unfamiliar, something which starts from the conceptual unintelligibility, in its present form, of the subjective-objective link. The enterprise is one of imagining possibilities: Identity theorists like Smart, Armstrong, and Lewis tried to explain how the identity of mental with physical states could be a contingent truth; I am interested in how some sort of mind-brain identity might be a necessary truth.

That would require not only the imagination of concepts that might capture the connection, but also some account of how our existing concepts would have to be related

to these and to one another. We must imagine something that falls under both our mental concepts and the physiological concepts used to describe the brain, not accidentally but necessarily.

IV. Mental Reference

We first have to interpret the third-person and first-person conditions of reference to mental states as inextricably connected in a single concept, but in a rather special way. I have insisted that mental concepts are not exhausted by the behavioral or functional conditions that provide the grounds for their application to others. Functionalism does not provide sufficient conditions for the mental. However in the other, "outward," direction there does seem to be a conceptual connection between conscious mental states and the behavioral or other interactions of the organism with its environment. This is a consequence of the inseparable first-person/third-person character of mental concepts. To put it roughly, functional states aren't necessarily mental states, but it is a conceptual truth that our mental states actually occupy certain functional roles.

Imaginability and thought experiments are essential in establishing conceptual connections--or their absence. Those methods have to be used with care, but the pitfalls are not so serious here as when they are used to test for nonconceptual necessary connections--as in the case of consciousness and the brain. We can discover the presence or absence of a conceptual connection a priori because all the necessary data are contained in the concepts we are thinking with: We just have to extract those data and see what they reveal.

Sometimes, as in the case of functional characteristics of consciousness, the conceptual connection may be somewhat hidden from view. But I believe we can know a priori both (a) that specific conscious states typically occupy certain functional roles, and (b) that those functional roles do not, as a matter of conceptual necessity, entail those specific conscious states. For the latter conclusion, we only have to imagine a being whose color vision, for example, is functionally equivalent to ours but is based on a completely different neurophysiology. This may not in fact be possible, but there is no reason to believe either that it is conceptually excluded, or that if it were possible, such a being would have the same color phenomenology as we do.

My main interest is in the further proposition that mental states are related to certain neurophysiological states by an equivalence relation that is necessary but not conceptual. But these other claims about the conceptual relation between phenomenology and behavior are an essential part of the picture. The aim is to connect phenomenology, physiology, and behavior in a single nexus.

I am denying two familiar types of functionalist view:

(1) Nonrigid functionalism: Mental concepts refer contingently to whatever inner states happen as a matter of fact to occupy certain functional roles. It is analytically true that to be a mental state of a given kind is simply to occupy a certain functional role, but it is contingently true of any particular inner state that it is a mental state of that kind. Empirical science reveals that mental concepts nonrigidly designate states that are in fact essentially physiological.[12]

[12] David Lewis, "Psychophysical and Theoretical Identifications," Australasian Journal

(2) Rigid functionalism: Mental concepts refer to functional states themselves--to the state of being in a state with a certain functional role. It is both analytically and necessarily true of a given mental state that it manifests itself in certain relations to behavior and to other mental states. Mental states are not identified with their physiological basis.[13]

The first view is unacceptable both because it analyzes mental concepts reductively and because it makes it a contingent fact that a mental state is the mental state it is. The second is unacceptable because it analyzes mental concepts reductively and implies that they don't refer to inner states of the organism.

Consider next the following alternative:

(3) Reference-fixing functionalism: The reference of our mental concepts to inner states is fixed by the contingent functional roles of those states, but the concepts apply rigidly to the occupants of those roles. It is neither necessarily true of a given mental state, nor analytically equivalent to its being the mental state it is, that it occupy a certain functional role, but that is how we in fact pick it out. Mental concepts rigidly designate states that are essentially physiological or phenomenological, or both.[14]

---

of Philosophy 1972

[13] Hilary Putnam, "The Nature of Mental States," in his Mind, Language and Reality: Philosophical Papers, vol. II (Cambridge University Press, 1975).

[14] This interesting option, which I had never heard before, was suggested to me by an

This seems to me close to the truth, but it leaves out the fact that the reference of mental terms for conscious states is fixed not only by their functional role but by their immediate phenomenological quality – an intrinsic and essential property. Something must relate these two reference-fixers, one necessary and one contingent, and I believe it can be done by the following proposal:

> (4) Though mental concepts cannot be analyzed functionally, functional roles are needed to fix the reference of mental terms, because of the inextricable first-person/third-person character of mental concepts. It is a conceptual but contingent truth that each mental state plays its characteristic functional role in relation to behavior. It is a conceptual and necessary truth that each conscious mental state has the phenomenological properties that it has. And it is a nonconceptual but necessary truth that each conscious mental state has the physiological properties that it has.

This seems to me to do justice to the "internal" character of the relation between phenomenology and behavior. Phenomenological facts have to be in principle, though not infallibly, introspectively accessible. If two simultaneous color impressions, or two sound impressions in close succession, are the same or different, I ought in general to be able to tell--just because they are both mine--and this discriminatory capacity will have

---

N.Y.U. undergraduate, James Swoyer. A theory of similar form, but offered in the serviceof physicalism, is defended by Michael E. Levin, Metaphysics and the Mind-Body Problem (Oxford: Clarendon Press, 1979), pp. 113-125.

behavioral consequences under suitable conditions. Similarly, if a sensation is very unpleasant, I will want to avoid it, and if I am not paralyzed this will also have behavioral consequences. Although phenomenological features cannot be analyzed behaviorally or functionally, their relation to their typical functional role in the production of behavior is, in the outward direction, an a priori conceptual truth.

This is the conception of the relation between mental states and behavior – conceptual but nonreductionist--that is suggested to me by Wittgenstein's anti-private-language argument, even though it is almost certainly not Wittgenstein's conception. If each phenomenal property were in principle detectable only introspectively, there could be no concepts for such properties, for the concepts could not be governed by rules that distinguished between their correct and incorrect application. Therefore our phenomenal concepts must actually work differently, picking out properties that are detectable from both the first-person and the third-person perspective. And this seems phenomenologically accurate, so long as it is not turned into a behaviorist or essentially third-person causal-role analysis of mental concepts. Pain, color impressions, and so forth are intrinsic properties of the conscious subject, which we can identify only in virtue of their relations to other mental properties and to causal conditions and behavioral manifestations.

To state Wittgenstein's point: In order to name a sensation that I notice, I must have the concept of the same (type of) sensation--of its feeling the same to me--and this must be the idea of something that can hold objectively, so that if I give the name "S" to the type of sensation I am having now, that baptism sets up a rule which determines

whether any particular future application of the term by me to another event will be correct or incorrect. It either will be the same--i.e. will feel the same phenomenally--or it will not. That I am correctly remembering the meaning of the term must be an objective fact independent of my actual sincere application of the term, or else the term wouldn't carry any meaning. So I must be relying on my mastery of a concept of phenomenal similarity to which my personal usage conforms over time--a concept whose applicability to me is independent of my application of it to myself, in a way that underwrites the objective meaning of my own personal application of it.

Concepts can be objective in more than one way, but phenomenological concepts seem in fact to secure their objectivity through an internal connection to behavior and circumstances. That is how we establish that someone else has the concept of sensation, and that is how an individual knows that he himself has mastered a phenomenological concept--by confirmation from others who can observe that he uses it correctly. It is also how we tell that we ourselves or someone else have forgotten what a phenomenological term means, and have misapplied it. The concept that we apply introspectively to ourselves is the same concept that others apply to us--and we to others--observationally.

To have the concept of pain a person must apply it to his own sensation in the circumstances that enable others to apply it to him. This conjunction is the only way to identify the concept. The third-person conditions are not sufficient, but they are (conceptually) necessary. Someone doesn't have the phenomenological concept of pain unless he can apply it introspectively in accordance with certain standard circumstantial and behavioral conditions. These include its tendency to signal damage and to provoke

avoidance, in an otherwise intact organism.

The reference of a phenomenological term is fixed, then, by its immediate phenomenological quality, whose identification depends on its functional role. A given functional role might be occupied by different phenomenological qualities in different organisms--or conceivably there could be a system in which the same functional role was not occupied by a conscious experience at all. And my hypothesis is that when a functional role helps to fix the reference of a sensation term, the term refers to something whose immediate phenomenological quality and physiological basis are both essential properties of it, properties without which it could not exist.

This is parallel to the case of water: There could be a watery liquid ("behaviorally" indistinguishable from water) that wasn't the compound H2O and therefore wasn't water; but in the world as it is, the essential gross properties of water are entailed by its being H2O, and that is what water is. Similarly, it is conceivable that there could be a state functionally equivalent to pain in a mechanism with a completely different internal constitution, and if it were both physically and phenomenologically different, it would not be the same sensation. But in us, the behavior that helps to fix the reference of "pain" is produced by a state whose phenomenological and physiological properties are both essential, and that is what pain is.

So the proposal is that mental states would have a dual essence – phenomenological and physiological--but we still don't understand how this could be, since our modal intuitions go against it. In particular, we still have to deal with the apparent conceivability of an exact chemical-physiological-functional replica of a

conscious human being that nevertheless has no subjective phenomenological "interior" at all--a zombie, in current jargon. This is an illusion, according to the above proposal, but it still has to be dissolved. The task of defending a necessary connection between the physical and the phenomenological requires some account of how a connection that is in fact internal remains stubbornly external from the point of view of our understanding.

Colin McGinn gives a similar description of the situation in his essay, "The Hidden Structure of Consciousness," though he puts it in terms of the distinction between the "surface" of consciousness and its true nature, inaccessible to us either by introspection or by external observation:

> My position is that the hidden structure of consciousness contains the machinery to lock consciousness firmly onto the physical world of brain and behaviour and environment, but that the surface of consciousness encourages us to believe that these links are merely contingent. When you cannot perceive (or conceive) necessary links you are apt to think there are not any, especially when you have racked your brains trying to discover them. This is a mistake, but a natural one. Cognitive closure with respect to necessary links is misinterpreted as contingency in those links.[15]

By "the surface of consciousness" I take him to mean the way it appears from the first-person standpoint--whether we are experiencing our own or imagining someone else's. This seems to be both something we have a very clear grasp of and something logically quite unconnected with the physical workings of the brain, even though there

[15] In The Problem of Consciousness (Blackwell, 1991), pp. 106-7, fn. 23.

are obviously causal connections. McGinn holds that both these appearances are illusory in a way we are prevented from seeing because we cannot get beneath the surface of consciousness.

V. What's Wrong with the Conceivability Argument

Though I believe McGinn is right about our present situation, I think we can advance beyond it once we acknowledge that our immediate first-person grasp on the phenomenology may be logically incomplete. But is that a real possibility? Perhaps our concepts of consciousness and the brain, while not containing full information about these two types of thing, are still adequate to allow us to know a priori that no necessary relation between them can be discovered no matter how much more we learn about their deeper constitutions. Perhaps the difference in type is such as to set limits on the paths along which fuller knowledge of the nature of these things can develop.

This is what seems forced on us by the clarity with which we appear to be able to conceive absolutely any physiological process existing unaccompanied by conscious experience. The vivid imaginability of a totally unconscious zombie, resembling a conscious being only in its behavior and physical constitution, seems not to depend in any way on the details of that constitution. That is because conceiving that the system has no consciousness is completely independent of conceiving anything about its physical character. The latter is a conception of it from the outside, so to speak, as a spatiotemporal structure, whereas the former is a conception of it from the subjective point of view, as having no subjective "inside" at all. The two types of conception are so

completely unrelated that the first seems incapable of ruling out the second: All I have to do is imagine the physical system from the outside, and then imagine it from the inside--as not having any inside in the experiential sense. That is, I project my own point of view into the zombie, and imagine that there is nothing of that kind going on behind its eyes at all. What could be more clearly independent than these two conceptions?[16]

But it is just the radical difference between these modes of conceiving that may undermine the result. I want now to argue not directly for the necessary connection between mind and brain, but rather for the position that even if there were such a necessary connection, it would still appear through this kind of conceivability test that there was not. The process of juxtaposing these two very different kinds of conception is inherently misleading.

In testing philosophical hypotheses by thought experiments, one should be wary of intuitions based on the first-person perspective, since they can easily create illusions of conceivability.[17] The zombie thought experiment clearly depends on the first-person perspective, because although it is an intuition about a being other than oneself, it

---

[16] This argument has recently been given much prominence by David Chalmers; see The Conscious Mind (Oxford University Press, 1996). It was thinking about Chalmers's book that stimulated me to write the present essay. And while we come to very different conclusions, there is a great deal in his book with which I agree.

[17] See Sydney Shoemaker, "The First-Person Perspective," in The First-Person Perspective and Other Essays (Cambridge University Press, 1996).

depends on taking up that being's point of view in imagination--or rather, finding that it has no point of view that one can take up. In this case the very disparity between the two forms of conception that gives rise to the strong intuition of conceivability should make us suspicious. The absence of any conceptual connection when phenomena are grasped by such disparate concepts may conceal a deeper necessary connection that is not yet conceptual because not accessible to us by means of our present forms of thought.

To see this, consider how I might investigate reflectively the relations among phenomenology, behavior, and physiology with respect to the taste of the cigar I am now smoking. What I must do first is to regard the experience as a state of myself of whose subjective qualities I am immediately aware, which also has certain publicly observable functional relations to stimuli and discriminatory capacities. Even at this first stage there is already the risk of a natural illusion of conceptual independence with respect to these functional relations, because they are concealed in my introspective identification of the experience. But it is an illusion because introspective identification is itself one of those mental acts that cannot be completely separated from its functional connections (for example the capacity to distinguish this taste from that of a cigarette). Recognizing this, I can see that the Cartesian thought-experiment of imagining myself having this experience without ever having had a body at all is an unreliable guide to what might really be the case. It depends on the concealment of the necessary conditions of reference of the phenomenological concept that I am employing to think about the experience. That is the point I take from Wittgenstein.

But now what of the relation between the experience and its physiological basis?

Here I seem to be able to imagine either myself or someone else tasting exactly this flavor of cigar--and its having all the usual functional connections as well--although my brain or the other person's brain is in a completely different physiological state from the one it is actually in. Indeed it seems imaginable, though unlikely, that when I offer a friendly cigar to an exotic visitor from outer space who has a completely different physiology, it should taste the same to him. But here too the imagination is a poor guide to possibility, because it relies on an assumption of the completeness of the manifest conditions of reference of the concept (now taken to include functional as well as phenomenological conditions).

The first thing to acknowledge is that if there were a necessary connection between the phenomenology and the physiology of tasting a cigar, it would not be evident a priori on the basis of the ordinary concept of that experience, since the possession of that concept involves no awareness of anything about the brain. It isn't just that, like the criterial connections of mental concepts to typical behavior and circumstances, the relation to the brain is hidden from view in my first-person use of the concept: The relation is completely absent from the concept, and cannot be retrieved by philosophical analysis. Nevertheless, if there is such a relation, having the full concept (including the first person aspect) would require having a brain, indeed a brain with exactly the right physiological characteristics, and the brain would be directly involved in the act of imagination--though its involvement would be completely outside the range of my awareness in employing the concept. To imagine a mental state from the inside would be what I have elsewhere called an act of sympathetic imagination--putting myself in a

conscious state resembling the thing imagined--and it would be impossible to do this without putting my brain in a corresponding physical state.[18]

This shows that I cannot rely on the apparent imaginability of the separation of phenomenology and physiology to establish the contingency of the relation, since I can know in advance that this act of imagination would seem subjectively the same whether the relation was contingent or necessary. If the relation is necessary, then I have not really succeeded in imagining the phenomenology without the physiology. The imagination here is essentially ostensive, and I cannot point to one without pointing to the other.

If the relation is necessary, then someone is mistaken if he says, concentrating on his present sensation of tasting a cigar, "I can conceive of this experience existing while my brain is in a very different state." He is mistaken because he is actually referring, by "this experience," to something that is at the same time a specific brain state. And if the relation is necessary, then someone is also mistaken who says, "I can conceive of the brain state that is in fact the physical condition of my tasting the cigar as existing without any such sensation existing." He is mistaken because he is actually referring, by "the brain state...," to something that is at the same time the experience. He does not really

---

[18] See "What Is It Like to Be a Bat?" (Philosophical Review, 1974; reprinted in Mortal Questions, Cambridge University Press, 1979) fn. 11. This was an earlier response to the modal argument against materialism. See also Christopher Hill, "Imaginability, Conceivability, Possibility and the Mind-Body Problem," Philosophical Studies, 1995.

succeed in detaching the one from the other in imagination, because he cannot demonstratively pick out either of them separately--even though the lack of visible connection between the two ways of picking out the same thing conceals this from him.

This does not show that the relation is necessary, but it does show that the familiar subjective thought experiment doesn't prove that the relation is contingent. The thought experiment would come out the same way whether the relation was necessary or contingent.

I think we can still rely on such thought experiments to refute the most common types of conceptual reductionism. Even if there is some kind of entailment of the mental by the physical-functional, it is not analytic or definitional: There is no hidden conceptual contradiction in the description of a zombie--even if in reality a zombie is logically impossible. Our mental concepts do not, for example, exclude the possibility that mental states are states of an immaterial soul, and that there could be a fully functioning physical replica of a human body without a soul. As I have said, this does not rule out a conceptual link in the other direction--from the mental to the behavioral – on account of the public criteria for the application of mental concepts, which go with their distinctive first-person/third-person character. But while third-person criteria are necessary for the operation of mental concepts, they are not sufficient. In any case, those criteria are functional rather than physiological, and the issue here is the relation between mental states and the brain, not between mental states and behavior. Here there is obviously no conceptual connection, and this tempts us to think that their separation is conceivable. But the inference is unwarranted.

The following things seem prima facie conceivable which are pretty certainly impossible in a very strong sense, namely:

(1) a living, behaving, physiologically and functionally perfect human organism that is nevertheless completely lacking in consciousness, i.e. a zombie;

(2) a conscious subject with an inner mental life just like ours that behaves and looks just like a human being but has electronic circuitry instead of brains.

The apparent conceivability of these things reveals something about our present concepts but not about what is really possible. Analytic psychophysical reductionism is false, but there is independent reason to believe that these are not logical possibilities, and if so, our concepts are missing something. They don't lead to contradiction – it's not as bad as that--but they fail to reveal a logical impossibility.

Contrast these thought experiments with the a priori inconceivability of a number having parents. The latter involves a straightforward clash between concepts, not merely a disparity. No number could enter into the kind of biological relation with a predecessor that is a necessary condition of being a child or offspring. In that case we see a contradiction between the conditions of numberhood and the conditions of being the child of anything or anyone. In the relation of consciousness to the physical world, by contrast, our concepts fail to reveal a necessary connection, and we are tempted to conclude to the absence of any such connection. Our intuition is of a logical compatibility, not of a logical incompatibility. We conceive the body from outside and the mind from inside, and see no internal connection, only an external one of correlation or perhaps causation.

Conceivability and inconceivability are the main evidence we have for possibility and necessity, but they can be misleading, and conceivability that depends on the relation between first and third person reference is particularly treacherous terrain. The first-person view of our experiential states may reveal something that is not just contingently related to their physical basis, despite appearances. The physical description of the brain states associated with consciousness may be an incomplete account of their essence--merely the outside view of what we recognize from within as conscious experience. If anything like that is true, then our present conceptions of mind and body are radically inadequate to the reality, and do not provide us with adequate tools for determining whether the relation between them is necessary or contingent.

VI. A New Concept

How am I to form the conception that the relation might actually be necessary--as opposed to merely acknowledging that I can't discover a priori that it isn't? I have to think that these two ways of referring--by the phenomenological concept and the physiological concept--pick out a single referent, in each case rigidly, but that the logical link cannot be discovered by inspecting the concepts directly: Rather it goes only through their common necessary link to the referent itself.

The idea would have to be, then, that there is a single event to which I can refer in two ways, both of them via concepts that apply to it noncontingently. One is the mental concept that I am able to acquire in both first and third person applications because I am a subject of this state, which has the special character of consciousness and introspective

accessibility--the state of tasting a cigar. The other is a (so far unspecified) physiological concept that describes the relevant physical state of the brain. To admit the possibility of a necessary connection here, we would have to recognize that the mental concept as it now operates has nothing to say about the physiological conditions for its own operation, and then open up the concept to amplification by leaving a place for such a condition--a place that can be filled only a posteriori, by a theory of the actual type of event that admits these two types of access, internal and external, from within and from without. But this description of the task tells us nothing about how to carry it out.

What will be the point of view, so to speak, of such a theory? If we could arrive at it, it would render transparent the relation between mental and physical, not directly, but through the transparency of their common relation to something that is not merely either of them. Neither the mental nor the physical point of view will do for this purpose. The mental will not do because it simply leaves out the physiology, and has no room for it. The physical will not do because while it includes the behavioral and functional manifestations of the mental, this doesn't enable it, in view of the falsity of conceptual reductionism, to reach to the mental concepts themselves. The right point of view would be one which, contrary to present conceptual possibilities, included both subjectivity and spatiotemporal structure from the outset, all its descriptions implying both these things at once, so that it would describe inner states and their functional relations to behavior and to one another from the phenomenological inside and the physiological outside simultaneously--not in parallel. The mental and physiological concepts and their reference to this same inner phenomenon would then be seen as secondary and each

partial in its grasp of the phenomenon: Each would be seen as referring to something that extends beyond its grounds of application.

Such a viewpoint cannot be constructed by the mere conjunction of the mental and the physical. It must be something genuinely new, otherwise it will not possess the necessary unity. It would have to be a new theoretical construction, realist in intention, and contextually defined as part of a theory that explained both the familiarly observable phenomenological and the physiological characteristics of these inner events. Its character would be determined by what it was introduced to explain--like the electromagnetic field, gravity, the atomic nucleus, or any other theoretical postulate. This could only be done with a truly general theory, containing real laws and not just dispositional definitions, otherwise the theoretical entity would not have independent reality.

If strict correlations are observed between a phenomenological and a physiological variable, the hypothesis would be not that the physiological state causes the phenomenological, but that there is a third term that entails both of them, but that is not defined as the mere conjunction of the other two. It would have to be a third type of variable, whose relation to the other two was not causal but constitutive. This third term should not leave anything out. It would have to be an X such that X's being a sensation and X's being a brain state both follow from the nature of X itself, independent of its relation to anything else.

Even though no transparent and direct explanatory connection is possible between the physiological and the phenomenological, but only an empirically established

extensional correlation, we may hope and ought to try as part of a scientific theory of mind to form a third conception that does have direct transparently necessary connections with both the mental and the physical, and through which their actual necessary connection with one another can therefore become transparent to us. Such a conception will have to be created; we won't just find it lying around. A utopian dream, certainly: but all the great reductive successes in the history of science have depended on theoretical concepts, not natural ones--concepts whose whole justification is that they permit us to give reductive explanations.

But there is another objection – that such extravagance is unnecessary. Why wouldn't a theory be sufficient that systematically linked mental phenomena to their physical conditions without introducing any concepts of a new type? That is the approach favored by John Searle, who maintains that a purely empirical theory would enable us to see that mental states are higher-order physical states of the brain, caused by lower-order physiological states to which they are not reducible.[19] Searle, too, wants to avoid dualism without resorting to functionalist reductionism, but I don't think his way of doing it succeeds. The problem is that so long as the mental states remain characteristically subjective and radically emergent, there is no basis for describing them as physical, or physically constituted.

This is not just a verbal point. The mental-physical distinction cannot be abolished by fiat. I agree with Searle that the correct approach to the mind-body problem must be essentially biological, not functional or computational. But his proposal is still,

[19] See The Rediscovery of the Mind (MIT Press, 1992).

as I understand it, too dualistic: In relating the physiological and the mental as cause and effect, it does not explain how each is literally impossible without the other. A causal theory of radically emergent higher-order properties would not show how mind arises from matter by necessity. That is the price of sticking with our existing mental and physical concepts.

The inadequacy of those concepts is revealed by their incapacity to display a necessary connection that obviously must exist. Only new concepts that turn the connection into a conceptual one can claim to grasp the phenomena in their basic nature.

Clearly not just any concept that we can create, which has both mental and physical implications, would reveal a necessary connection between the two. In some cases, we will only have created a conjunctive concept, relative to which the two categories are analytically, but not necessarily, connected. For example, even if Cartesian dualism were true, we could introduce the concept of a human being as the combination of a body and a soul. In that case there would be one thing, a human being, whose existence entails both mental and physical characteristics, but that would not mean that one can't exist without the other, any more than the concept of a ham sandwich shows that bread can't exist without ham.

What is the difference between these purely conjunctive, analytic connections, and the more metaphysically robust type of concept that reveals true necessity? Physical science is full of examples of the latter. The clearest are found in the atomic theory of matter. The hypothesis that familiar substances are composed of invisibly small particles, whose motion is responsible for the observable manifestations of temperature and

pressure, made it possible to see that the positive correlation between changes in temperature and pressure of a gas at constant volume was not a contingent but a necessary connection. Likewise the chemical analysis of air, and of fire as rapid oxidation, reveals it to be a necessary truth that a fire will go out if enclosed in a small airtight space. The postulation of electromagnetic fields, similarly, made it possible to see many previously mysterious correlations, such as the capacity of a moving magnet to induce an electric current, as necessary consequences of the nature of the component phenomena--though in this case the new concept requires a greater leap from prescientific intuition than the direct analogy with the familiar part-whole relation that yields atomism.

One of the things that is true in these cases is that the "single" postulated underlying phenomenon explains the manifestations of each of the superficially distinct phenomena in a way that makes it impossible to separate the explanation of the one from the explanation of the other. The very same atomic (or molecular) agitation that accounts for increased pressure against the walls of the container accounts for increased temperature of the gas within. The process of oxidation that constitutes the fire eventually binds all the free oxygen in the airtight container, thus entailing its own termination. So the new account of the correlated phenomena makes their separability no longer conceivable.[20]

[20] Given the character of modern physics, all these necessities have to understood probabilistically.

In addition, the postulated underlying basis explains more things than it was introduced to explain. Atomic theory was the avenue to the endless developments of chemistry; the theory of electromagnetism led vastly beyond the curious phenomena of lodestones and electrostatic charge from which it began. It is clear that such postulates cannot be analyzed in terms of the manifestations on the basis of which they were introduced, since they imply so much more that is not implied by those manifestations themselves. For all these reasons, the unification accomplished by such concept formation is not merely verbal, or conjunctive. It is the genuine discovery that things that appeared distinct and only contingently correlated are in fact, in virtue of their true nature, necessarily connected.

So the discovery of a genuinely unifying, rather than conjunctive, basis for the relation between mind and body would require the postulation of something that accounted for them both in terms of the same activity, or properties, or structure, or whatever. And its reality would be confirmed if it could also account for things other than those it had been postulated to explain or their direct implications--other, previously unremarked psychophysical correlations, for example. That would require more than an inference from observed correlations to psychophysical laws that in turn predict further correlations. It would mean finding something that entailed such laws as the logically necessary consequence of its essential nature.

It is a real question whether there is something already present in our current concepts of mental and physical--some unbridgeable gulf--that precludes their both being accounted for in the requisite unified way by a common basis. The atomistic method, of

accounting for a property of the whole by explaining all its physical manifestations in terms of the activities of the parts, is not sufficient here, because there is more to be explained than the observable physical manifestations of mental processes.

Merely adding phenomenological qualities to brain states as an extra property is not enough, since it would imply that the same brain state might exist without that property. It has to follow from what these states really are that they have both these types of properties. If we are going to take reduction in physics and chemistry as our logical model, we have to recognize, as was explained earlier, that the necessary identity of water with H2O or fire with oxidation or heat with molecular motion depends on another necessary connection. It requires that the manifest properties by which we prescientifically identify water or fire or heat must be explained without residue, and in their essential respects entailed, by the reducing account. This upward entailment--that all the distinguishing marks of heat are in fact exhaustively explained by molecular motion--is essential for the validity of the downward entailment--that heat is identical with molecular motion and cannot exist without it. The only way we can discover that heat is molecular motion--so that if something felt the same to us but was not molecular motion it would not be heat--is to discover that in our world the actual complete account of the features by which we identify heat pretheoretically is given in terms of molecular motion, and that this account is complete in the sense that it entails what is essential in those features.

In the mind-body case, there is no direct entailment in either direction between the phenomenological and the physiological, and at present we don't have the concept of

a third type of state or process that would entail both the phenomenological and the physiological features of an experiential episode like tasting chocolate. But that is what would be required to warrant the conclusion that tasting chocolate had this physiological character necessarily, or vice versa. Only if we discovered such an actual common basis would we be able to say that a zombie is impossible, as water that is not H2O is impossible, or fire that is not oxidation.

If we did discover such a thing, it would perhaps still be conceivable that something should look outwardly like a living human being with a functioning brain but not have consciousness. But such a system would have to be constituted out of different material, and would therefore not, despite appearances, be a physical duplicate of a human body, merely lacking consciousness. On the supposition that in us, the psychophysical connection is necessary, the brain of such a creature could not be made of what our brains are made of, and would be similar only in its external appearance--just as there might be a different colorless, odorless, tasteless liquid that was not H2O and therefore not water.21

---

21 This leaves us with a further question. Suppose we did discover such a common basis. Would there not then be an analogue, for the zombie case, of the possibility of another liquid that resembles water in its manifest qualities, but that is not water because it is not H2O? Can we imagine something like that with respect to consciousness and the brain?

The question can be divided in two parts. First, even if our conscious states were in fact brain states, couldn't we imagine a different physical system that to external observation resembled a human being in all behavioral and physiological and chemical respects, but

VII. Under Mind

I have described these conditions for the existence of a necessary connection between phenomenology and physiology very abstractly. They do not yet offer any suggestion of what kind of concept might entail both, and thus reveal their common foundation. It would have to be the concept of something that in its essence has, and cannot fail to have, both a subjective inside and a physical outside.

Let me at last, after this very long windup, offer an extravagant conjecture. I suggest that we take the macroscopic relation between mental processes and their

---

consisted of intrinsically different material that lacked consciousness? Second, couldn't we imagine a different conscious subject with experience that subjectively resembled human pain, but that was not pain (!) because it was not a brain state but, say, a state of an immaterial soul?

I believe that in both of these cases, unlike the water case, there is no reason to think that we have imagined any possibility at all. Even if such alternative systems were possible, our use of our own imagination of the presence or absence of subjective experience could give us no evidence of it. If the connection between our minds and our brains is indeed necessary, then our imagination provides no way of peeling off the experience from its physical embodiment, or vice versa, as argued in the previous section. We have no way of conceiving of the presence or absence of the purely mental features of experience by themselves. By contrast, we do have a way of conceiving of the presence or absence of the perceptual appearances of water by themselves, since those appearances involve a relation to something else, namely the perceiver.

behavioral manifestations, which I have said is conceptual but not necessary, as a rough model for a deeper psychophysical connection that is necessary – pushing embodiment inward, so to speak. The gross and manifest relations between consciousness and behavior would thus be reinterpreted as a rough indicator of something much tighter in the interior of the brain, that can be discovered only by scientific inference, and that explains the manifest relations in virtue of its usual links to the rest of the body. Perhaps, for example, the reason for the relation between pain and avoidance at the level of the organism is that at a deeper physiological level, the state that generates the appropriate observable behavior in an intact organism by the mediation of nerves, muscles, and tendons is an essentially subjective state of the brain with an unmediated, noncontingent "behavioral expression" of its own. It would be a single state that is necessarily both physical and mental, not a mere conjunction of the two.

Does this "pushing down" of the relation between mind and its behavioral manifestations make sense: Could there be a tighter version of the relation below the level of the whole organism? Well, to begin with the first level down, these relations should certainly be reflected in some form in the case of a separated but still operating brain--the classic imaginary "brain in a vat"--deprived of its body but still receiving inputs and producing outputs, and functioning internally otherwise like an embodied brain. Its mental states (I assume it would have mental states) would bear a relation to its purely electronic inputs and outputs analogous to those of a normal person to perceptual inputs and behavioral outputs – but without the contingency due to dependence on the usual external connections.

The next question is whether the same is true of half brains. In the case of individuals with brain damage, or those with split brains, each functioning cerebral hemisphere seems to interact with the brain stem in a way that expresses behaviorally the somewhat reduced conscious activity associated with the partial brain. I believe the remarkable split brain results have a philosophical significance that has not been sufficiently appreciated.[22] They show that both the brain and the mind are in some sense composed of parts, and that those parts are simultaneously physical and mental systems, which can to some extent preserve their dual nature when separated. In an intact brain, the two halves do not lead distinct conscious lives: They support a single consciousness. But the fact that each of them can support a distinct consciousness when separated seems to show that the normal unified consciousness is composed of mental parts embodied in the physical parts. These parts are "mental" in a derivative but nonetheless real sense.

If this phenomenon of composition can be seen to exist at the gross level of bisection, it makes sense to conjecture that it may be carried further, and that some form of more limited psychophysical unity may exist in smaller or more specialized subparts of the system, which in ordinary circumstances combine to form a conscious being of the familiar kind, but may also in some cases be capable of existing and functioning separately. The strategy would be to try to push down into the interior of the brain the supposition of states loosely resembling ordinary mental states in that they combine

---

[22] I have discussed those results in "Brain Bisection and the Unity of Consciousness," (Synthese, 1971; reprinted in Mortal Questions, Cambridge University Press, 1979).

constituents of subjective mental character (in an extended sense) with behavioral or functional manifestations – with the difference that here the "behavior" would be internal to the brain, rather than being mediated by links to the body--an intrinsic, noncontingent feature of the state rather than a relation to something outside of it. And they need not be spatially defined subparts, but might include other sorts of subsystems or operations that are not strictly localized.

Such hypothetical subparts of consciousness would not be subjectively imaginable to us. They would be subjective only in the sense that they are inherently capable of combining to constitute full states of consciousness in an intact organism, even though they have no independent consciousness when they are so combined, and may or may not have independent consciousness when they occur separately. The compositional character of consciousness is evident not only from bisected brains but from the description of people with the sort of brain damage that causes behaviorally spectacular and subjectively alien mental changes. Certain cases of agnosia are like this, as when a person can pick a pen out of a group of objects if asked to do so, but can't say, if shown a pen, what it is, and can't show how it is used--though he can when he touches it. This is due to some cut between the visual, tactile, and speech centers, and it isn't really imaginable from the inside to those who don't suffer from it.[23]

A theory of the basis of the mental-physical link might begin from the component

[23] See Norman Geschwind, "Disconnexion Syndromes in Animals and Man," Brain 88 (1965) for extensive discussion of such disorders.

analysis suggested by the deconnection syndromes. Some such pushing down of the link to a level lower than that of the person is necessary to get beyond brute emergence or supervenience. Even if crude spatial divisions are only part of the story, they might be a beginning. More global but functionally specialized psychophysical subsytems might follow. The conceptual point is that both the mind and the brain may be composed of the same subsystems, which are essentially both physical and mental, and some variants of which are to be found in other conscious organisms as well.

The idea of a third type of phenomenon--essentially both mental and physical--which is the real nature of these subprocesses is easier to grasp if one thinks of the mental aspect as irreducibly real but not subjectively imaginable from an ordinary complete human viewpoint. It would be conceivable only by inference from what can be observed – inference precisely to what is needed to explain the observations. Constituents inferred to explain simultaneously both the physiological and the phenomenological data and the connection between them would not be classifiable in the old style either as physical or as mental. We would have to regard the physical results of combining such constituents in a living organism—results we could observe both behaviorally and physiologically--as providing only a partial view of them.

Such a compositional theory would be one possible way and perhaps the only way to give content to the idea of a necessary connection between the physiological and the mental. To me it seems clear that any necessary connection must be a matter of detail, and not just global. The necessary connection between two things as complex as a creature's total mental state and its total physiological state must be a consequence of

something more fundamental and systematic. We can't form the conception of a necessary connection in such a case just by stipulating that they are both essential features of a single state. The inseparability must be the logical consequence of something simpler to avoid being a mere constant conjunction that provides evidence of necessity without revealing it. Necessity requires reduction, because in order to see the necessity we have to trace it down to a level where the explanatory properties are simply the defining characteristics of certain basic constituents of the world.

Our ordinary sensation concepts paint these states with a broad brush. We all know that in our own case there is much more detail, both phenomenological and physiological, than we can describe in ordinary language. The systematic though imprecise relation at the level of the organism between mind and behavior is captured by ordinary mental concepts, but it is only the rough and macroscopic manifestation of objective lawlike conditions that must lie much deeper. And the detailed macroscopic relation between mind and brain may be necessary, though it appears contingent, because it is the consequence of the noncontingent physiological manifestations of component states at a submental level.

This hypothesis invites several questions. First, would the states I am imagining at the basic level really be unified, rather than raising again the question of the relation between their mental and physical aspects? Second, can we really make sense of the idea of each mind being composed of submental parts? Third, what is the relation between the physicality of these submental processes and the account of what happens in the brain in terms of physics and chemistry alone?

The first question requires us to distinguish a manifestation of a property that is truly essential, revealing an internal, noncontingent relation, from one that is due to a merely contingent, external relation.

All our working concepts require that there be some form of generally available access to what they refer to, and that means that any concept of a type of process or substance, or of a property, mental or physical, will refer to something that is systematically connected to other things, allowing different people from their different points of view to get a handle on it. This is the grain of truth in verificationism. It is true whether the property is liquidity or heat or painfulness. There are no natural kinds without systematic connections to other natural kinds.[24] All properties that we can think about have to be embedded in a web of connections, and I suspect that this is even true of properties we can't think about, because it is part of our general concept of a property.

Sometimes the properties that permit us to make contact with a natural kind are external, contingent properties. This, I have said, is true of the ordinary behavioral manifestations of mental states, that permit us to have public mental concepts. It is also true of the manifest properties by which we fix the reference of many other natural kind

---

[24] This position is much more fully and precisely expressed and defended by Sydney Shoemaker in his remarkable paper, "Causality and Properties," in Identity, Cause and Mind: Philosophical Essays (Cambridge University Press, 1984), with which I agree entirely. He also points out that it is a consequence of this view, fully worked out, that causal necessity is a species of logical necessity.

terms. But the closer we get to the thing itself, the more unmediated will be its manifestations, its effects, and its relations to certain other things. Eventually we arrive at effects that are directly entailed by the essential properties of the natural kind itself. The mass and charge of a proton, for example, without which it would not be a proton, have strict consequences for its relations to other particles, similarly specified. Even in describing radically counterfactual situations we have to suppose these essential relations preserved in order to be sure we are talking about the same property or thing. Some dispositions are necessary consequences of a thing's essential nature.

Let us look more closely at the familiar physical case. The manifest properties of ordinary physical objects--their shape, size, weight, color, and texture, for example--already have necessary consequences for their interactions with other things whose properties are specified with sufficient precision. The properties are not reducible to those external relations, but the consequences are not merely contingent. An object simply would not weigh one pound if it did not affect a scale in the appropriate way, in the absence of countervailing forces. But all these necessary connections at the gross level have implications for the type of analysis at the level of physical theory that can reveal more fully the intrinsic nature of such an object. An analysis in terms of microscopic components, however strange and sophisticated its form, must in some way preserve these necessary external relations of the properties of the manifest object. The properties of the parts may be different--a crude mechanistic atomism, while a natural presocratic speculation, has proved too simple--but they must have their own necessary consequences for interaction with other things, of a kind that in combination will imply

the relational properties of the larger entity which they compose. However far we get from the manifest world of perception and common sense, that link must not be broken. Even if some of the properties of the whole are emergent in the sense of not being predictable from the separately ascertainable properties of the parts, the emergent phenomena still consist of or are constituted by the collective behavior of the parts.

Something similar is needed if our starting point is not the manifest world of inanimate physical objects, but the world of conscious creatures. In a case like thirst, for example, the subjective quality and the functional role are already internally connected in the ordinary concept. It is the concept of a phenomenological state that has typical physical manifestations. The full intrinsic character of the state has to be discovered. But the ordinary concept already contains, in rough form, an idea of the kind of state it is – just as an ordinary substance concept like water already contains, in rough form, an idea of the kind of thing it is, setting the possible paths to further detailed discovery of its true nature, which have led to the development of physics and chemistry.

The hypothesis of psychic atoms that are just like animals, only smaller, is not even a starter in this case, because we don't have ready a coherent idea of larger conscious subjects being composed of smaller ones--as the early atomists had the perfectly clear geometrical idea of larger physical objects or processes being composed of smaller ones. But the more abstract idea of a form of analysis of conscious organisms whose elements will preserve in stricter form the relation between mental reality and behavior should constrain and guide the development of any reductive theory in this domain. There must be some kind of strict inner-outer link at more basic levels that can

account for the far looser and more complicated inner-outer link at the level of the organism. And of course the idea would have to include a completely new theory of composition--of mental parts and wholes. (As I have said, the parts and wholes would include not just chunks of the brain and their smaller components, but nonspatially defined processes and functions as well.)

My conjecture is that the relation between conscious states and behavior, roughly captured in the way ordinary mental concepts function, is a manifest but superficial and contingent version of the truth—namely, that the active brain is the scene of a system of subpersonal processes which combine to constitute both its total behavioral and its phenomenological character, and that each of those subpersonal processes is itself a version of a "mental-behavioral" relation that is not contingent but necessary because it is not mediated by anything.

This differs from traditional functionalism, coupled with an account of the physiological realization of functional states, in that the "realization" here envisioned is to be not merely physiological, but in some sense mental all the way down--something that accounts for the phenomenology as well. The combination of these postulated processes would entail at more complex levels not only the observable behavior and functional organization but the conscious mental life conceptually related but not reducible to it. We are looking for a realization not just of functional states but of mental states in the full sense, and that means the realization cannot be merely physical. The reductive basis must preserve, in broad terms, the logical character of the mental processes being reduced. That is just as true here as it is in reductions of purely physical

substances, processes, or forces.

The problem of adequate unity in the inferred explanatory concept--the problem of how it can avoid being a mere conjunction of the phenomenological and the physiological--can be addressed by seeing it as a purification of the ordinary concept of mind, with the sources of contingency in the mental-behavioral connections gradually removed as we close in on the thing itself. States of this kind, if they exist, could be identified only by theoretical inference; they would not be definable as the conjunction of independently identifiable mental and physical components, but would be understandable only as part of a theory that explains the relations between them.

I leave aside the question of how far down these states might go. Perhaps they are emergent, relative to the properties of atoms or molecules. If so, this view would imply that what emerges are states that are in themselves necessarily both physical and mental—not just mental states attached to nonemergent physical states. If, on the other hand, they re not emergent, this view would imply that the fundamental constituents of the world, out of which everything is composed, are neither physical nor mental, but something more basic. This position is not equivalent to panpsychism. Panpsychism is, in effect, dualism all the way down.25 This is monism all the way down.

I said there were three questions about the proposal. The second was how we could conceive of a single mind resulting from the combination of subpersonal components. On that issue, we have very few data to go on, only the split brain cases. Further experiments to investigate the results of combining parts of different conscious

25 See “Panpsychism,” in Mortal Questions (Cambridge University Press, 1979).

nervous systems would be criminal if carried out on human subjects--the only kind who would be able to tell us about the experiential results. (There's a piece of science fiction for you.) But the contents of an animal mind are complex enough so that the idea of composition seems a fairly natural one--though who knows what kinds of "parts" the combinable components might be. We certainly can't expect them in general to be anatomically separable. The now common habit of thinking in terms of mental modules is a crude beginning, but it might lead somewhere, and might join naturally with the creation of concepts of the sort I am suggesting, which entail both physiology and phenomenology. The real conceptual problems would come in trying to describe elements or factors of subjectivity too basic to be found as identifiable parts of conscious experience. I will not try to say more about compositionality at this point.

The third question was about the relation between explanation employing such concepts and such a theory on the one hand, and traditional, purely physical explanation on the other. The idea is that such a theory would explain both the phenomenology and the physiology by reference to a more fundamental level at which their internal relation to one another was revealed. But wouldn't that require that there be no account of the physical interactions of a conscious organism with its environment, and of its internal physical operation, in terms of the laws of physics and chemistry alone? Whether or not such an account is possible, the denial of its possibility would certainly seem a dangerously strong claim to harness to any hypothesis of the kind I am suggesting.

My quick response to this question is that there is no reason to think that the explanations referring to this psychophysical level need conflict with purely physical

explanations of the purely physical features of the same phenomena, any more than explanations in terms of physics have to conflict with explanations in terms of chemistry. If there is a type of description which entails both the mental and the physical, it can be used to explain more than what a purely physical theory can explain, but it should also leave intact those explanations that need to refer only to the physical. If there are special problems here, they have to do with the compatibility between psychological and physical explanations of action, and freedom of the will. Those problems are serious, but they are not, I think, made any more serious by a proposal of this kind, whereby the relation between the mental and the physiological is necessary rather than contingent. Indeed, such a proposal would probably dispose of one problem, that of double causation, since it would imply that at a deeper level the distinction between mental and physical causes disappears.

VIII. Universal Mind

All this is speculation of the most extravagant kind, but not for that reason impermissible. Armchair proto-science as the philosophical formulation of possibilities is an indispensable precondition of empirical science, and with regard to the mind-body problem we are not exactly awash in viable possibilities.[26] I have described in abstract

[26] See "Philosophical Naturalism," Michael Friedman's presidential address to the Central Division of the American Philosophical Association, Proceedings and Addresses of the American Philosophical Association, vol. 71 no. 2 (November 1997), pp. 7-21.

terms the logical character of a different theory and different concepts. Their creation, if possible at all, would have to be based on empirical research and theoretical invention. But one feature such a theory should have that is of the first importance is a universality that extends to all species of conscious life, and is not limited to the human. That just seems to me to be common sense about how the world works. The mind-brain relation in us must be an example of something quite general, and any account of it must be part of a more general theory. That conception ought to govern us even if we have to start with humans and creatures very like them in gathering evidence on which to base such a theory.

This has an important consequence for the basic theoretical terms it will employ, the terms which entail both the phenomenological and the physiological descriptions of inner states. They must be understood to imply that experiences have a subjective character, without necessarily allowing the theorist to fully understand the specific subjective character of the experiences in question – since those experiences may be of a type that he himself cannot undergo or imagine, and of which he cannot therefore acquire the full first- and third-person mental concepts. The terms will therefore have to rely, in their full generality, a good deal on what I have elsewhere[27] called "objective phenomenology"--structural features like quality spaces that can be understood and described as aspects of a type of subjective point of view without being fully subjectively imaginable except by those who can share that point of view.

If such a theory is ever developed, the reason for believing in the reality of what it

[27] In "What Is It Like to Be a Bat?"

postulates, like the reason for believing in the reality of any other theoretical entities, will be inference to the best explanation. The relation between phenomenology and physiology demands an explanation; no explanation of sufficient transparency can be constructed within the circle of current mental and physical concepts themselves; so an explanation must be sought which introduces new concepts and gives us knowledge of real things we didn't know about before. We hypothesize that there are things having the character necessary to provide an adequate explanation of the data, and their real existence is better confirmed the wider the range of data the hypothesis can account for. But they must be hypothesized as an explanation of the mental and the physical data taken together, for there will be no reason to infer them from physiological and behavioral data alone. As Jeffrey Gray observes,

> The reason the problem posed by consciousness seems so acute, at least to nonfunctionalists, is the following: nothing that we know so far about behaviour, physiology, the evolution of either behaviour or physiology, or the possibilities of constructing automata to carry out complex forms of behaviour is such that the hypothesis of consciousness would arise if it did not occur in addition as a datum in our own experience; nor, having arisen, does it provide a useful explanation of the phenomena observed in those domains.[28]

The most radical thing about the present conjecture is the idea that there is

---

[28] Jeffrey A. Gray, "The contents of consciousness: A neuropsychological conjecture," Behavioral and Brain Sciences (1995) 18, p. 660.

something more fundamental than the physical--something that explains both the physical and the mental. How can the physical be explained by anything but the physical? And don't we have ample evidence that all that needs to be postulated to get ever deeper explanations of physical phenomena is just more physics? However I am not proposing that we look for a theory that would displace or conflict with physical explanation of the ordinary sort--any more than it would conflict with ordinary psychological explanation of actions or mental events. Clearly the processes and entities postulated by such a theory would have to conform to physical law. It's just that there would be more to them than that. What reveals itself to external observation as the physiological operation of the brain, in conformity with physical law, would be seen to be something of which the physical characteristics were one manifestation and the mental characteristics another--one being the manifestation to outer sense and the other the manifestation to inner sense, to adapt Kant's terminology.

This leaves open the question of the level and type of organization at which the stuff becomes not just dead matter but actually conscious: Its mental potentialities might be completely inert in all but very special circumstances. Still, it would have to explain the mental where it appears, and in a way that also explains the systematic connections between the mental and the physical and the coexistence of mental and physical explanations, as in the cases of thought and action. And this conception would, if it were correct, provide a fuller account of the intrinsic nature of the brain than either a phenomenological or a physiological description, or the conjunction of the two.

To describe the logical characteristics of such a theory is not to produce it. That

would require the postulation of specific theoretical structures defined in terms of the laws governing their physical and mental implications, experimentally testable and based on sufficiently precise knowledge of the extensional correlations between physical and mental phenomena. The path into such a theory would presumably involve the discovery of systematic structural similarities between physiological and phenomenological processes, leading eventually to the idea of a single structure that is both, and it would have to be based on vastly more empirical information than we have now.

It would have to be graspable by us, and therefore would have to be formulated in terms of a model that we could work with, to accommodate psychophysical data that we do not yet have. But it would not be simply an extension of our existing ideas of mind and matter, because those ideas do not contain within themselves the possibility of a development through which they "meet."

I have suggested one possible form of an approach that would permit such convergence, but it would not permit us to transcend the division between subjective and objective standpoints. The aim is rather to integrate them all the way to the bottom of our world view, in such a way that neither is subordinate to the other. This means that what Bernard Williams calls the "absolute" conception of reality[29] will not be a physical conception, but something richer that entails both the physical and the mental. To the extent that we could arrive at it, it would describe subjective experience in general terms that imply its subjectivity without necessarily relying on our capacity to undergo or fully imagine experiences of that type. That means that our grasp of such an absolute

[29] Bernard Williams, Descartes (Penguin Books, 1978).

conception will inevitably be incomplete. Still, it would include more than a purely physical description of reality.

Whatever unification of subjective points of view and complex physical structures may be achieved, each of us will still be himself, and will conceive of other perspectives by means of sympathetic imagination as far as that can reach, and by extrapolation from imagination beyond that. The difference between the inside and the outside view will not disappear. For each of us, the site and origin of his conception of the world as a unified physical-phenomenological system will always be the particular creature that he himself is, and therefore the conception will have a centered shape that is at variance with its centerless content. But that need not prevent us from developing that content in a way that captures the evident unity of what in our own case we can experience both from within and from without.

Previous efforts at reduction have been too external and in a sense too conservative. We need a conceptual creation that, by revealing a hidden necessary connection, makes conceivable what at present is inconceivable, so it won't be possible to imagine such a theory properly in advance. But it won't be possible even to look for such a solution unless we start with an incomplete conception of it. And that requires the willingness to contemplate the idea of a single natural phenomenon that is in itself, and necessarily, both subjectively mental from the inside and objectively physical from the

outside—just as we are. [30]

[30] Some portions of this essay derive from "Conceiving the Impossible and the Mind-Body Problem," Philosophy 73 (1998), pp. 337-352.