PHILOSOPHY OF MIND

THIRD EDITION

JAEGWON KIM

WESTVIEW
PRESS

A Member of the Perseus Books Group

Westview Press was founded in 1975 in Boulder, Colorado, by notable publisher and intellectual Fred Praeger. Westview Press continues to publish scholarly titles and high-quality undergraduate- and graduate-level textbooks in core social science disciplines. With books developed, written, and edited with the needs of serious nonfiction readers, professors, and students in mind, Westview Press honors its long history of publishing books that matter.

Published by Westview Press,
A Member of the Perseus Books Group

Find us on the World Wide Web at www.westviewpress.com.

Designed by Trish Wilkinson
Set in 10.5 point Minion Pro

Library of Congress Cataloging-in-Publication Data

Kim, Jaegwon.
Philosophy of mind / Jaegwon Kim.—3rd ed.
p. cm.
ISBN 978-0-8133-4458-4 (alk. paper)
1. Philosophy of mind. I. Title.
BD418.3.K54 2011
128'.2—dc22
E-book ISBN 978-0-8133-4520-8 2010040944

10 9 8 7 6 5 4 3 2 1

CHAPTER 6

Mind as a Causal System

Causal-Theoretical Functionalism

In the preceding chapter, we discussed the functionalist attempt to use Turing machines to explicate the nature of mentality and its relationship to the physical. Here we examine another formulation of functionalism, in terms of "causal role." Central to any version of functionalism is the idea that a mental state can be characterized in terms of the input-output relations it causally mediates, where the inputs and outputs may include other mental states as well as sensory stimuli and physical behaviors. Mental phenomena are conceived as nodes in a complex causal network that engages in causal transactions with the outside world at its peripheries, by receiving sensory inputs and emitting behavior outputs.

What, according to functionalism, distinguishes one mental kind (say, pain) from another (say, itch) is the distinctive input-output relationship associated with each kind. Causal-theoretical functionalism conceives of this input-output relationship as a causal relation, one that is mediated by mental states. Different mental states are different because they are implicated in different input-output causal relationships. Pain differs from itch in that each has its own distinctive causal role: Pains typically are caused by tissue damage and cause winces, groans, and escape behavior; in contrast, itches typically are caused by skin irritation and cause scratching. But tissue damage causes pain only if certain other conditions are present, some of which are mental in their own right; not only must you have a properly functioning nervous system, but you must also be normally alert and not engrossed in another task. Moreover, among the typical effects of pain are further mental states, such as a feeling of distress and a desire to be relieved of it. But this seems to involve us in a regress or circularity: To explain what a given mental state is, we need to refer to other mental states, and explaining these can only be expected to require reference to further mental states, and so on—a process that can go on in an unending regress or loop back in a circle. Circularity threatens to arise at a more general level as well, in the functionalist conception of mentality itself: To be a mental state is to be an internal state serving as a causal intermediary between sensory inputs and *mental states* as causes, on the one hand, and behaviors and other *mental states* as effects, on the other. Viewed as a definition of what it is to be a mental state, this is obviously circular. To circumvent the threatened circularity, machine functionalism exploits the concept of a Turing machine in characterizing mentality. To achieve the same end, causal-theoretical functionalism exploits the entire network of causal relations involving all psychological states—in effect, a comprehensive psychological theory—to anchor the physical-behavioral definitions of individual mental properties.[1]

The Ramsey-Lewis Method

Consider the following toy "pain theory":

> (T) For any *x*, if *x suffers tissue damage* and **is normally alert**, *x* **is in pain**; if *x is awake*, *x* tends to be **normally alert**; if *x* **is in pain**, *x winces* and *groans* and **goes into a state of distress**; and if *x* **is not normally alert** or *x* **is in a state of distress**, *x tends to make more typing errors*.

We assume that the statements constituting T describe lawful regularities (or causal relations). The italicized expressions are nonmental predicates designating observable physical, biological, and behavioral properties; the expressions in boldface are psychological predicates designating mental properties. T is, of course, much less than what we know about pain and its relationship to other events and states, but let us assume that T encapsulates what is important about our knowledge of pain. Issues about the kind of "theory" T must be if T is to serve as a basis of functional definitions of mental expressions will be taken up in a later section. Here T is only an example to illustrate the formal technique originally due to Frank P. Ramsey, a British mathematician-philosopher in the early twentieth century, and later adapted by David Lewis for formulating functional definitions of mental kinds.[2]

We first "Ramseify" T by "existentially generalizing" over each psychological predicate occurring in it, which yields this:

(T_R) There exist states M_1, M_2, and M_3 such that for any x, if x *suffers tissue damage* and is in M_1, x is in M_2; if x *is awake*, x tends to be in M_1; if x is in M_2, x *winces* and *groans* and goes into M_3; and if x is either not in M_1 or is in M_3, x *tends to make more typing errors*.

The main thing to notice about T_R vis-à-vis T is that instead of referring (as T does) to specific mental states, T_R speaks only of *there being some states or other*, M_1, M_2, and M_3, which are related to each other and to observable physical-behavioral states in the way specified by T. Evidently, T logically implies T_R (essentially in the manner in which "x is in pain" logically implies "There is some state M such that x is in M"). Note that in contrast to T, its Ramseification T_R contains no psychological expressions but only physical-behavioral expressions such as "suffers tissue damage," "winces," and so on. Terms like "M_1," "M_2," and "M_3" are called predicate variables (they are like the *x*s and *y*s in mathematics, though these are usually used as "individual" variables)—they are "topic-neutral" logical terms, neither physical nor psychological. Expressions like "is normally alert" and "is in pain" are predicate constants, that is, actual predicates.

Ramsey, who invented the procedure now called "Ramseification," showed that although T_R is weaker than T (since it is implied by, but does not imply, T), T_R is just as powerful as T as far as physical-behavioral prediction goes; the two theories make exactly the same deductive connections between nonpsychological statements.[3] For example, both theories entail that if someone is awake and suffers tissue damage, she will wince, and that if she does not groan, either she has not suffered tissue damage or she is not awake. Since T_R is free of psychological expressions, it can serve as a basis for defining psychological expressions without circularity.

To make our sample definitions manageable, we abbreviate T_R as "$\exists M_1, M_2, M_3[T(M_1, M_2, M_3)]$." (The symbol $\exists$, called the "existential quantifier," is read: "there exist.") Consider, then:[4]

x is in pain $=_{\text{def}} \exists M_1, M_2, M_3[T(M_1, M_2, M_3)$ and x is in $M_2]$

Note that "M_2" is the predicate variable that replaced "is in pain" in T. Similarly, we can define "is alert" and "is in distress" (although our little theory T was made up mainly to give us a reasonable definition of "pain"):

x is normally alert $=_{\text{def}} \exists M_1, M_2, M_3$ $[T(M_1, M_2, M_3)$ and x is in $M_1]$

x is in distress $=_{\text{def}} \exists M_1, M_2, M_3$ $[T(M_1, M_2, M_3)$ and x is in $M_3]$

Let us see what these definitions say. Consider the definition of "being in pain": It says that you are in pain just in case there are certain states, M_1, M_2, and M_3, that are related among themselves and with such physical-behavioral states as tissue damage, wincing and groaning, and typing performance as specified in T_R *and* you are in M_2. It is clear that this definition gives us a concept of pain in terms of its causal-nomological relations and that among its causes and effects are other "mental" states (although these are not specified as such but referred to only as "some" states of the psychological subject) as well as physical and behavioral events and states. Notice also that there is a sense in which the three mental concepts are interdefined but without circularity; each of the defined expressions is completely eliminable by its definiens (the right-hand side of the definition), which is completely free of psychological expressions. Whether or not these definitions are adequate in all respects, it is evident that the circularity problem has been solved.

So the trick is to define psychological concepts holistically en masse. Our T is a fragment of a theory, something made up to show how the method works; to generate more realistic functional definitions of psychological concepts by the Ramsey-Lewis method, we need a comprehensive underlying psychological theory encompassing many more psychological kinds and richer and more complex causal-nomological relationships to inputs and outputs. Such a theory will be analogous to a Turing machine that models a full psychology, and the resemblance of the present method with the approach of machine functionalism should be clear, at least in broad outlines. In fact, we can think of the Turing machine approach as a special case of the Ramsey-Lewis method in which the psychological theory is presented in the form of a Turing machine table with the internal machine states, the *q*s, corresponding to the predicate variables, the Ms. We discuss the relationship between the two approaches in more detail later.

Choosing an Underlying Psychology

So what should the underlying psychological theory T be like if it is to yield, by the Ramsey-Lewis technique, adequate functional definitions of psychological properties? If we are to recover a psychological property from T_R by the Ramsey-Lewis method, the property must appear in T to begin with. So T must refer to all psychological properties. Moreover, T must carry enough information about each psychological property—about how it is nomologically connected with input conditions, behavior outputs, and other psychological properties—to circumscribe it closely enough to identify it. Given this, there are two major possibilities to consider.

We might, with Lewis, consider using the platitudes of our shared *commonsense psychology* as the underlying theory. The statements making up our "pain theory" T are examples of such platitudes, and there are countless others about, for instance, what makes people angry and how angry people behave, how wants and beliefs combine to generate further wants, how perceptions cause beliefs and memories, and how beliefs lead to further beliefs. Few people are able to articulate these principles of "folk psychology," but most mature people use them constantly in attributing mental states to people, making predictions about how people will behave, and understanding why people do what they do. We know these psychological regularities "tacitly," perhaps in much the way we "know" the grammar of the language we speak—without being able to state any explicit rules. Without a suitably internalized commonsense psychology in this sense, we would hardly be able to manage our daily transactions with other people and enjoy the kind of communal life that we take for granted.[5] It is important that the vernacular psychology that serves as the underlying theory for functional definitions consists of *commonly known* generalizations. This is essential if we are to ensure that functional definitions yield the psychological concepts that all of us share. It is the shared funds of vernacular psychological knowledge that collectively define our commonsense mental concepts; there is no other conceivable source from which our mental concepts could magically spring. Functionalism that takes these psychological platitudes as a basis for functional definitions of psychological terms is sometimes called "analytical functionalism." The thought is that these well-known psychological generalizations are virtually "analytic" truths—truths that are evident to speakers who understand the meanings of the psychological expressions involved.

We must remember that commonsense psychology is, well, only commonsensical: It may be incomplete and partial, and contain serious errors, or even inconsistencies. If mental concepts are to be defined in terms of causal-nomological relations, shouldn't we use our best theory about how mental events and states are involved in causal-nomological relations, among themselves and with physical and behavioral events and processes? *Scientific psychology*, including cognitive science, after all, is in the business of investigating these regularities, and the best scientific psychology we can muster *is* the best overall theory about the causal-nomological facts of mental events and states. The form of functionalism that favors empirical scientific theory as the Ramseification base is sometimes called "psycho-functionalism."

There are problems and difficulties with each of these choices. Let us first note one important fact: If the underlying theory T is false, we cannot count on any mental concepts defined on its basis to apply to anything—as logicians will say, these concepts will have empty, or null, extensions.[6] For if T is false, its Ramseification, T_R, may also be false; in particular, if T has false nonmental consequences (for example, T makes wrong behavioral predictions), T_R will be false as well. (Recall that T and T_R have the same physical-behavioral content.) If T_R is false, every concept defined on its basis by the Ramsey-Lewis method will be vacuous—that is, it will not apply to anything. This is easy to see for our sample "pain theory" T. Suppose this theory is false—in particular, suppose that what T says about the state of distress is false and that in fact there is no state that is related, in the way specified by T for distress, with the other internal states and inputs and outputs. This makes our sample T_R false as well, since there is nothing that can fill in for M_3. This would mean that "pain" as defined on the basis of T_R cannot be true of anything: Nothing satisfies the defining condition of "pain." The same goes for "normally alert" and "the state of distress." So if T, the underlying theory, is false, all mental concepts defined on its basis by the Ramsey-Lewis method will turn out to have the same extension, namely, the null extension!

This means that we had better make sure that the underlying theory is true. If our T is to yield our psychological concepts all at once, it is going to be a long conjunction of myriad psychological generalizations, and even a single false component will make the whole conjunction false. So we must face these questions: What is going to be included in our T, and how certain can we be that T is true? Consider the case of scientific psychology: It is surely going to be a difficult, probably impossible, task to decide what parts of current scientific psychology are well enough established to be considered uncontroversially true. Psychology has been flourishing as a science for many decades now, but it is comparatively young as a science, with its methodological foundations still in dispute, and it is fair to say that it has yet to produce a robust enough common core of generally accepted laws and theories. In this respect, psychology has a long way to go before it reaches the status of, say, physics, chemistry, or even biology.

These reflections lead to the following thought: On the Ramsey-Lewis method of defining psychological concepts, every dispute about the underlying theory T is going to be a dispute about psychological concepts. This creates a seemingly paradoxical situation: If two psychologists should disagree about some psychological generalization that is part of theory T, which we should expect to be a common occurrence, this would mean that they are using different sets of psychological concepts. But this seems to imply that they cannot really disagree, since the very possibility of disagreement presupposes that the same concepts are shared. How could I accept and you reject a given proposition unless we shared the concepts in terms of which the proposition is formulated?

Perhaps things are not as bleak as they seem. For example, there is probably no need to invoke a total psychology as a base for functional definitions of mental terms; relatively independent parts of psychology and cognitive science, like theory of vision, theory of motivation, decision and action, theory of language acquisition, and so on, can each serve as a basis of Ramseification. Also we can consider degrees of similarity between concepts, and it may be possible for two speakers to understand each other well enough in a given situation, without sharing an exactly identical set of concepts; sharing similar concepts may be good enough for the purposes at hand.

Consider again the option of using commonsense psychology to anchor psychological concepts. Can we be sure that all of our psychological platitudes, or even any of them, are true—that is, that they hold up as systematic scientific psychology makes progress? Some have even argued that advances in scientific psychology have already shown commonsense psychology to be massively erroneous and that, considered as a theory, it must be abandoned.[7] Consider the generalization, used as part of our pain theory, that tissue damage causes pain in a normally alert person. It is clear that there are many exceptions to this regularity: A normally alert person who is totally absorbed in another task may not feel pain when she suffers minor tissue damage. Massive tissue damage may cause a person to go into a coma. And what is to count as "normally alert" in any case? Alert enough to experience pain when one is hurt? The platitudes of commonsense psychology may serve us competently enough in our daily life in anticipating behaviors of our fellow humans and making sense of them. But are we prepared to say that they are literally true? One way to alleviate these worries is to point out that we should think of our folk-psychological generalizations as hedged by generous escape clauses ("all other things being equal," "under normal conditions," "in the absence of interfering forces," and so on). Whether such weakened, noncommittal generalizations can introduce sufficiently restrictive constraints to yield well-defined psychological concepts is something to think about.

In one respect, though, commonsense psychology seems to have an advantage over scientific psychology: its apparently greater stability. Theories and concepts of systematic psychology come and go; given what we know about the rise and fall of scientific theories, especially in the social and human sciences, it is reasonable to expect that most of what we now consider our best theories in psychology will be abandoned and replaced, or seriously revised, sooner or later—probably sooner rather than later. The rough regularities codified in commonsense psychology appear considerably more stable (perhaps because they are rough); can we really imagine giving up the virtual truism that a person's desire for something and her belief that doing a certain thing will secure it tends to cause her to do it? This basic principle, which links belief and desire to action, is a central principle of commonsense psychology that makes it possible to understand why people do what they do. It seems reasonable to think that the principle was as central to the way the ancient Greeks or Chinese made sense of themselves and their fellows as it is to our own folk-psychological explanatory practices. Our shared folk-psychological heritage is what enables us to understand, and empathize with, the actions and emotions of the characters depicted in Greek tragedies and historical Chinese fiction. Indeed, if there were a culture, past or present, for whose members the central principles of our folk psychology, such as the one that relates belief and desire to action, did not hold true, its institutions and practices would hardly be intelligible to us, and its language might not even be translatable into our own. The source and nature of this relative permanence and commonality of folk-psychological platitudes are in need of explanation, but it seems plausible that folk psychology enjoys a degree of stability and universality that eludes scientific psychology.

We should note, though, that vernacular psychology and scientific psychology need not necessarily be thought to be in competition with each other. We could say that vernacular psychology is the appropriate underlying theory for the functional definition of vernacular psychological concepts, while scientific psychology is the appropriate one for scientific psychological concepts. If we believe, however, that scientific psychology shows, or has shown, vernacular psychology to be seriously flawed (for example, showing that many of its central generalizations are in fact false),[8] we would have to reject the utility of the concepts generated from it by the Ramsey-Lewis method, for as we saw, these concepts would then apply to nothing.

There is one final point about our sample functionalist definitions: They can accommodate the phenomenon of multiple realization of mental states. This is easily seen. Suppose that the original pain theory, T, is true of both humans and Martians, whose physiology, let us assume, is very different from ours (it is inorganic, say). Then T_R, too, would be true for both humans and Martians: It is only that the triple of physical-biological states <H_1, H_2, H_3>, which realizes the three mental states <pain, normal alertness, distress> and therefore satisfies T_R for humans, is different from the triple of physical states

<I_1, I_2, I_3>, which realizes the mental triple in Martians. But in either case there exists a triple of states that are connected in the specified ways, as T_R demands. So when you are in H_2, you are in pain, and when Mork the Martian is in I_2, he is in pain, since each of you satisfies the functionalist definition of pain as stated.

Functionalism as Physicalism: Psychological Reality

Let us return to scientific psychology as the underlying theory to be Ramseified. As we noted, we want this theory to be a true theory. Now, there is another question about the truth of psychological theories that we need to attend to. Let us assume that psychological theories posit internal states to systematize correlations between sensory inputs and behavioral outputs. These internal states are the putative psychological states of the organism. Suppose now that each of two theories, T_1 and T_2, gives a correct systematization of inputs and outputs for a psychological subject S, but that each posits a different set of internal states. That is, T_1 and T_2 are both *behaviorally adequate* psychologies for S, but each attributes to S a different internal causal structure connecting S's inputs to its outputs. Is there some further fact about these theories, or about S, that determines which (if any) is the correct psychology of S? As a basis for Ramseified functional definitions of mental states, we presumably must choose the correct psychology *if* there is a correct one.

If psychology is a truly autonomous special science, under no methodological, theoretical, or metaphysical constraints from any other science, we would have to say that the only ground for preferring one or the other of two behaviorally adequate theories consists in broad formal considerations of notational simplicity, ease of deriving predictions, and the like. There could be no further fact-based grounds favoring one theory over the other. As you will recall, behaviorally adequate psychologies for subject S are analogous to Turing machines that are "behavioral descriptions" of S (see chapter 5). You will also recall that according to machine functionalism, not every behavioral description of S is a correct psychology of S and that a correct psychology is one that is a machine description of S—namely, a Turing machine that is physically realized by S. This means that there are internal physical states of S that realize the internal machine states of the Turing machine in question—that is, there are in S "real" internal physical states that are (causally) related to each other and to sensory inputs and behavioral outputs as specified by the machine table of the Turing machine. It is the requirement of physical realization that answers the question of the psychological reality of Turing machines purporting to specify the psychologies of a subject.

Unlike machine functionalism, causal-theoretical functionalism, formulated on the Ramsey-Lewis model, does not as yet have a physical requirement built into it. According to machine functionalism as formulated in the preceding chapter, for subject S to be in any mental state, S must be a *physical realization* of an appropriate Turing machine; in contrast, causal-theoretical functionalism as developed thus far in this chapter requires only that there be "internal states" of S that are connected among themselves and to inputs and outputs as specified by S's psychology, without saying anything about the nature of these internal states. What we saw in connection with machine functionalism was that it is the further physical requirement—to the effect that the states of S that realize the machine's internal states be physical states—that makes it possible to pick out S's correct psychology. In the same way, the only way to settle the issue of psychological reality between behaviorally adequate psychologies is to explicitly introduce a similar physicalist requirement, perhaps something like this:

> (P) The states that the Ramseified psychological theory, T_R, affirms to exist are physical-neural states; that is, the variables M_1, M_2, . . . of T_R and in the definitions of specific mental states (see our sample definitions of "pain," and so on) range over physical-neural states of the subjects of psychological theory T.

A functionalist who accepts (P) may be called a physicalist functionalist. Unless some physical constraints, represented by (P), are introduced, there seems to be no way of discriminating between behaviorally adequate psychologies. Conversely, the apparent fact that we do not think all behaviorally adequate psychologies are "correct" or "true" signifies our commitment to the reality of the internal, theoretical states posited by our psychologies, and the only way this psychological realism is cashed out is to regard these states as internal *physical* states of the organism involved. This is equivalent in substance to the thesis of realization physicalism discussed in the preceding chapter—the thesis that all psychological states, if realized, must be physically realized.

This appears to reflect the actual research strategies in psychology and cognitive science and the methodological assumptions that undergird them: The correct psychological theory must, in addition to being behaviorally adequate, have "physical reality" in the sense that the psychological capacities, dispositions, and mechanisms it posits have a physical (presumably neurobiological) basis. The psychology that gives the most elegant and simplest systematization of human behavior may not be the true psychology, any more than the simplest artificial intelligence program (or Turing machine) that accomplishes a certain intelligent task (proving logic theorems, face recognition, or whatever) accurately reflects the way we humans perform it. The psychological theory that is formally the most elegant may not describe the way humans (or other organisms or systems under consideration) actually process their sensory inputs and produce behavioral outputs. There is no reason, either a priori or empirical, to believe that the mechanism that underlies our psychology, something that has evolved over many millions of years in the midst of myriad unpredictable natural forces, must be in accord with our notion of what is simple and elegant in a scientific theory. The psychological capacities and mechanisms posited by a true psychological theory must be real, and the only reality to which we can appeal in this context seems to be physical reality. These considerations, quite apart from the arguments pro and con concerning the physical reducibility of psychology, cast serious doubts on the claim that psychology is an autonomous science not answerable to lower-level physical-biological sciences.

The antiphysicalist might argue that psychological capacities and mechanisms have their own separate, nonphysical reality. But it is difficult to imagine what they could be when dissociated from their physical underpinnings; could they be some ghostly mechanisms in Cartesian mental substances? That may be a logically possible position, but hardly a plausible one, philosophically or scientifically (see chapter 2). It isn't for nothing that physicalism is the default position in contemporary philosophy of mind and psychology.

Objections and Difficulties

In this section, we review several points that are often thought to present major obstacles to the functionalist program. Some of the problematic features of machine functionalism discussed in the preceding chapter apply, mutatis mutandis, to causal-role functionalism, and these will not be taken up again here.

Qualia Inversion

Consider the question: What do all instances of pain have in common in virtue of which they are pains? You will recognize the functionalist answer: their characteristic causal role—their typical causes (tissue damage, trauma) and effects (pain behavior). But isn't there a more obvious answer? What all instances of pain have in common in virtue of which they are all cases of pain is that they *hurt*. Pains hurt, itches itch, tickles tickle. Is there anything more obvious than that?

Sensations have characteristic *qualitative* features; these are called "phenomenal" or "phenomenological" or "sensory" qualities; "qualia" ("quale" for singular) is now the standard term. Seeing a ripe tomato has a certain distinctive visual quality that is unmistakably different from the visual quality involved in seeing a mound of spinach leaves. We are familiar with the smells of roses and ammonia; we can tell the sound of a drum from that of a gong; the feel of a cool, smooth granite countertop as we run our fingers over it is distinctively different from the feel of sandpaper. Our waking life is a continuous feast of qualia—colors, smells, sounds, and all the rest. When we temporarily lose our ability to taste or smell properly because of a bad cold, eating a favorite food can be like chewing cardboard and we are made acutely aware of what is missing from our experience.

By identifying sensory events with causal roles mediating input and output, however, functionalism appears to miss their qualitative aspects altogether. For it seems quite possible that causal roles and phenomenal qualities come apart, and the possibility of "qualia inversion" seems to prove it. It would seem that the following situation is perfectly coherent to imagine: When you look at a ripe tomato, your color experience is like my color experience when I look at a bunch of spinach, and vice versa. That is, your experience of red might be qualitatively like my experience of green, and your experience of green is like my experience of red. These differences need not show up in any observable behavioral differences: We both say "red" when we are shown ripe tomatoes, and we both describe the color of spinach as "green"; we are equally good at picking tomatoes out of mounds of lettuce leaves; and when we drive, we cope equally well with the traffic lights. In fact, we can coherently imagine that your color spectrum is systematically inverted with respect to mine, without this being manifested in any behavioral differences. Moreover, it seems possible to think of a system, like an electromechanical robot, that is functionally—that is, in terms of inputs and outputs—equivalent to us but to which we have no good reason to attribute any qualitative experiences (again, think of Commander Data; this is called the "absent qualia" problem).[9] If inverted qualia, or absent qualia, are possible in functionally equivalent systems, qualia cannot be captured by functional definitions, and functionalism cannot be an account of all psychological states and properties. This is the qualia argument against functionalism.

Can the functionalist offer the following reply? On the functionalist account, mental states are realized by the internal physical states of the psychological subject; so for humans, the experience of red, as a mental state, is realized by a specific neural state. This means that you and I cannot differ in respect of the qualia we experience as long as we are in the same neural state; given that both you and I are in the same neural state, something that is in principle ascertainable by observation, either both of us experience red or neither does.

But this reply falls short for two reasons. First, even if it is correct as far as it goes, it does not address the qualia issue for physically different systems (say, you and the Martian) that realize the same psychology. Nothing it says makes qualia inversion impossible for you and the Martian; nor does it rule out the possibility that qualia are absent from the Martian experience. Second, the reply assumes that qualia supervene on the physical-neural states that realize them, but this supervenience assumption is part of what is at issue. However, the issue about qualia supervenience concerns the broader issues about physicalism; it is not specifically a problem with functionalism.

This issue concerning qualia has been controversial, with some philosophers doubting the coherence of the very idea of inverted or absent qualia.[10] We return to the issue of qualia in connection with the more general questions about consciousness (chapters 9 and 10).

The Cross-Wired Brain

Let us consider the following very simple, idealized model of how pain and itch mechanisms work: Each of us has a “pain box” and an “itch box” in our brains. We can think of the pain box as consisting of a bundle of neural fibers somewhere in the brain that gets activated when we experience pain, and similarly for the itch box. When pain receptors in our tissues are stimulated, they send neural signals up the pain input channel to the pain box, which then gets activated and sends signals down its output channel to our motor systems to cause appropriate pain behavior (winces and groans). The itch mechanism works similarly: When a mosquito bites you, your itch receptors send electrochemical signals up the itch input channel to your itch box, and so on, finally culminating in your itch behavior (scratching).

Suppose that a mad neurosurgeon rewires your brain by crisscrossing both the input and output channels of your pain and itch centers. That is, the signals from your pain receptors now go to your (former) itch box and the signals from this box now trigger your motor system to emit winces and groans; similarly, the signals from your itch receptors are now routed to your (former) pain box, which sends its signals to the motor system, causing scratching behavior. And suppose that I escape the mad neurosurgeon’s attention. It is clear that even though your brain is cross-wired with respect to mine, we both realize the same functional psychology: We both scratch when bitten by mosquitoes, and wince and groan when our fingers are burned. From the functionalist point of view, we instantiate the same pain-itch psychology.

Suppose that we both step barefoot on an upright thumbtack; both of us give out a sharp shriek of pain and hobble to the nearest chair. I am in pain. But what about you? The functionalist says that you, with the cross-wired brain, are in pain also. What makes a neural mechanism inside the brain a pain box is exactly the fact that it receives input from pain receptors and sends output to cause pain behavior. With the cross-wiring of your brain, your former itch box has now become your pain box, and when it is activated, you are in pain. At least that is what the functionalist conception of pain implies. But is this an acceptable consequence?

This is a version of the inverted qualia problem: Here the qualia that are inverted are pain and itch (or the painfulness of pains and the itchiness of itches), where the supposed inversion is made to happen through anatomical intervention. Many will feel a strong pull toward the thought that if your brain has been cross-wired as described, what you experience when you step on an upright thumbtack is an itch, not a pain, in spite of the fact that the input-output relation that you exhibit is one that is appropriate for pain. The appeal of this hypothesis is, at bottom, the appeal of the psychoneural identity theory of mentality. Most of us have a strong, if not overwhelming, inclination to think that types of conscious experience, such as pain and itch, supervene on the *local* states and processes of the brain no matter how they are hooked up with the rest of the body or the external world, and that the qualitative character of our mental states is conceptually and causally independent of their causal roles in relation to sensory inputs and behavioral outputs. Such an assumption is implicit, for example, in the popular philosophical thought-experiment with “the brain in a vat,” in which a brain detached from a human body is kept alive in a vat of liquid and maintained in a normal state of consciousness by being fed electric signals generated by a supercomputer. The qualia we experience are causally dependent on the inputs: As our neural system is presently wired, cuts and pinpricks cause pains, not itches. But this is a contingent fact about our neural circuitry: It seems perfectly conceivable (even technically feasible at some point in the future) to reroute the causal chains involved so that cuts and pinpricks cause itches, not pains, and skin irritations cause pains, not itches, without disturbing the overall functional organization of our behavior.

Functional Properties, Disjunctive Properties, and Causal Powers

The functionalist claim is often expressed by assertions like "Mental states are causal roles" and "Mental properties (kinds) are functional properties (kinds)." We should get clear about the logic and ontology of such claims. The concept of a functional property and related concepts were introduced in the preceding chapter, but let us briefly review them before we go on with some difficulties and puzzles for functionalism. Begin with the example of pain: For something, S, to be in pain (that is, for S to have, or instantiate, the property of being in pain) is, according to functionalism, for S to be in some state (or to instantiate some property) with causal connections to appropriate inputs (for example, tissue damage, trauma) and outputs (pain behavior). For simplicity, let us talk uniformly in terms of *properties* rather than *states*. We may then say: The property of being in pain is the property of having some property with a certain causal specification, in terms of its causal relations to certain inputs and outputs. Thus, in general, we have the following canonical expression for all mental properties:

> Mental property M is the property of having a property with causal specification H.

As a rule, the functionalist believes in the multiple realizability of mental properties: For every mental property M, there will in general be multiple properties, Q_1, Q_2, . . . , each meeting the causal specification H, and an object will count as instantiating M just in case it instantiates one or another of these Qs. As you may recall, a property defined the way M is defined is often called a "second-order" property; in contrast, the Qs, their realizers, are "first-order" properties. (No special meaning needs to be attached to the terms "first-order" and "second-order"; these are relative terms—the Qs might themselves be second-order relative to another set of properties.) If M is pain, then its first-order realizers are neural properties, at least for organisms, and we expect them to vary across various pain-capable biological species.

This construal of mental properties as second-order properties seems to create some puzzles. If M is the property of having some property meeting specification H, where Q_1, Q_2, . . . , are the properties satisfying H—that is, the Qs are the realizers of M—it would seem to follow that M is identical with the *disjunctive* property of having Q_1 or Q_2 or . . . Isn't it evident that to have M just *is* to have either Q_1 or Q_2 or . . . ? (For example, red, green, and blue are primary colors. Suppose something has a primary color; doesn't that amount simply to having red or green or blue?) Most philosophers who believe in the multiple realizability of mental properties deny that mental properties are disjunctive properties—disjunctions of their realizers—for the reason that the first-order realizing properties are extremely diverse and heterogeneous, so much so that their disjunction cannot be considered a well-behaved property with the kind of systematic unity required for propertyhood. As you may recall, the rejection of such disjunctions as legitimate properties was at the heart of the multiple realization argument against psychoneural-type physicalism. Functionalists have often touted the phenomenon of multiple realization as a basis for the claim that the properties studied by cognitive science are formal and abstract—abstracted from the material compositional details of the cognitive systems. What our considerations appear to show is that cognitive science properties so conceived threaten to turn out to be heterogeneous disjunctions of properties after all. And these disjunctions seem not to be suitable as nomological properties—properties in terms of which laws and causal explanations can be formulated. If this is right, it would disqualify mental properties, construed as second-order properties, as serious scientific properties.

But the functionalist may stand her ground, refusing to identify second-order properties with the disjunctions of their realizers, and she may reject disjunctive properties in general as bona-fide properties, on the ground that from the fact that both P and Q are properties, it does not follow that there is a disjunctive property, that of having P or Q. From the fact that being round and being green are properties, it does not follow, some have argued, that there is such a property as being round or green; some things that have this "property" (say, a red round table and a green square doormat) have nothing in common in virtue of having it. However, we need not embroil ourselves in this dispute about disjunctive properties, for the issue here is independent of the question about disjunctive properties.

For there is another line of argument, based on broad causal considerations, that seems to lead to the same conclusion. It is a widely accepted assumption, or at least a desideratum, that mental properties have causal powers: Instantiating a mental property can, and does, cause other events to occur (that is, cause other properties to be instantiated). In fact, this is the founding premise of causal-theoretical functionalism. Unless mental properties have causal powers, there would be little point in worrying about them. The possibility of invoking mental events in explaining behavior, or any other events, would be lost if mental properties should turn out to be causally impotent. But on the functionalist account of mental properties, just where does a mental property get its causal powers? In particular, what is the relationship between mental property M's causal powers and the causal powers of its realizers, the Qs?

It is difficult to imagine that M's causal powers could magically materialize on their own; it is much more plausible to think—it probably is the only plausible thing to think—that M's causal powers arise out of those of its realizers, the Qs. In fact, not only do they "arise out" of them, but the causal powers of any given instance of M must be the same as those of the particular Q*i* that realizes M on that occasion. Carburetors can have no causal powers beyond those of the physical device that performs the specified function of carburetors, and an individual carburetor's causal powers must be exactly those of the particular physical device in which it is realized (if for no other reason than the simple fact that this physical device *is* the carburetor).[11] To believe that it could have excess causal powers beyond those of the physical realizer is to believe in magic: Where *could* they possibly come from? And to believe that the carburetor has fewer causal powers than the particular physical device realizing it seems totally unmotivated; just ask, "Which causal powers should we subtract?"

Let us consider this issue in some detail. On functionalism, for a psychological subject to be in mental state M is for it to be in a physical state P where P realizes M—that is, P is a physical state that is causally connected in appropriate ways with other internal physical states (some of which realize other mental states) and physical inputs and outputs. In this situation, all that there is, when the system is in mental state M, is its physical state P; being in M has no excess reality over and beyond being in P, and whatever causal powers that accrue to the system in virtue of being in M must be those of state P. It seems evident that this instance of M can have no causal powers over and beyond those of P. If my pain, here and now, is realized in a particular event of my C-fibers being stimulated, the pain must have exactly the causal powers of the particular instance of C-fiber stimulation.

But we must remember that M is multiply realized—say, by P_1, P_2, and P_3 (the finitude assumption will make no difference). If multiplicity has any meaning here, these Ps must be importantly different, and the differences that matter must be *causal* differences. To put it another way, the physical realizers of M count as different because they have different, even extremely diverse, causal powers. For this reason, it is not possible to associate a unique set of causal powers with M; each *instance* of M, of course, is an instance of P_1 or an instance of P_2 or an instance of P_3 and as such represents a specific set of causal powers, those associated with P_1, P_2, or P_3. However, M taken as a kind or property does not. That is to say, two arbitrary M-instances cannot be counted on to have much in common in their causal powers beyond the functional causal role that defines M. In view of this, it is difficult to regard M as a property with any causal-nomological unity, and we are led to think that M has little chance of entering into significant lawful relationships with other properties. All this makes the scientific usefulness of M highly problematic.

Moreover, it has been suggested that kinds in science are individuated on the basis of causal powers; that is, to be recognized as a useful property in a scientific theory, a property must possess (or be) a determinate set of causal powers.[12] In other words, the resemblance that defines kinds in science is primarily *causal-nomological resemblance*: Things that are similar in causal powers and play similar roles in laws are classified as falling under the same kind. Such a principle of individuation for scientific kinds would seem to disqualify M and other multiply realizable properties as scientific kinds. This surely makes the science of the Ms, namely, the psychological and cognitive sciences, a dubious prospect.

These are somewhat surprising conclusions, not the least because most functionalists are ardent champions of psychology and cognitive science—in fact, of all the special sciences—as forming irreducible and autonomous domains in relation to the underlying physical-biological sciences, and this is the most influential and widely received view concerning the nature and status of psychology. We should remember that functionalism itself was largely motivated by the recognition of the multiple realizability of mental properties and a desire to protect the autonomy of psychology as a special science. The ironic thing is that if our reasoning here is not entirely off target, the conjunction of functionalism and the multiple realizability of the mental leads to the conclusion that psychology is in danger of losing its unity and integrity as a science. On functionalism, then, mental kinds are in danger of fragmenting into their multiply diverse physical realizers and ending up without the kind of causal-nomological unity and integrity expected of scientific kinds.[13]

Roles Versus Realizers: The Status of Cognitive Science

Some will object to the considerations that have led to these deflationary conclusions about the scientific status of psychological and cognitive properties and kinds as functionally conceived. Most functionalists, including many practicing cognitive and behavioral scientists, will find them surprising and unwelcome. For they believe, or want to believe, all of the following four theses: (1)

psychological-cognitive properties are multiply realizable; hence, (2) they are irreducible to physical properties; however, (3) this does not affect their status as legitimate scientific kinds; from all this it follows that (4) the cognitive and behavioral sciences form an autonomous science irreducible to more basic, "lower-level" sciences like biology and physics.

The defenders of this sort of autonomy thesis for cognitive-behavioral science will argue that the alleged fragmentation of psychological-cognitive properties as scientific properties, presented in the preceding section, was made plausible by our single-minded focus on their lower-level realizers. It is this narrow focus on the diversity of the possible realizers of mental properties that makes us lose sight of their unity as properties—the kind of unity that is invisible "bottom up." Instead, our focus should be on the "roles" that define these properties, and we should never forget that psychological-cognitive properties are "role" properties. So we might want to distinguish between "role functionalism" and "realizer functionalism."[14] Role functionalism identifies each mental property with being in a state that plays a specified causal role and keeps them clearly distinct from the physical mechanisms that fill the role, that is, the mechanisms that enable systems with the mental property to do what they are supposed to do. In contrast, realizer functionalism associates mental properties more closely with their realizers and identifies each specific *instance* of a mental property with an *instance* of its physical realizer. So the different outlooks of the two functionalisms may be stated like this:

> *Realizer Functionalism.* My experiencing pain at time *t* is identical with my C-fibers being activated at *t* (where C-fiber activation is the pain realizer in me); the octopus's experiencing pain at *t* is identical with its X-fibers being activated at *t* (where X-fiber activation is the octopus's pain realizer); and so on. The property instantiated when I experience pain at *t* is not identical with the property instantiated by the octopus when it experiences pain at *t*.

> *Role Functionalism.* My experiencing pain at time *t* is identical with my being at *t* in a state that plays causal role R (that is, the role of detecting bodily damage and triggering appropriate behavioral responses); the octopus's experiencing pain at *t* is identical with its being, at *t*, in a state that plays the same causal role R; and so on. My pain at *t* and the octopus pain at *t* share the same functional property, namely being in a state with causal role R.

Where the realizer functionalist sees differences and disunity among instances of pain, the role functionalist sees similarity and unity represented by pain's functional role. The role property associated with being in pain is what all pains have in common, and the role functionalist claims that these role properties are thought to constitute the subject matter of psychology and cognitive science; the aim of these sciences is to discover laws and regularities holding for these properties, and this can be done without attending to the physical and compositional details of their realizing mechanisms. In this sense, these sciences operate with entities and properties that are abstracted from the details of the lower-level sciences. Going back to the four theses, (1) through (4), it will be claimed that they should be understood as concerning mental properties as conceived in accordance with role functionalism.

Evidently, for role properties to serve these purposes, they must be robustly causal and nomological properties. Here is what Don Ross and David Spurrett, advocates of role functionalism, say:

> The foundational assumptions of cognitive science, along with those of other special sciences, deeply depend on role functionalism. Such functionalism is crucially supposed to deliver a kind of causal understanding. Indeed, the very point of functionalism (on role *or* realizer versions) is to capture what is salient about what systems actually do, and how they interact, *without* having to get bogged down in micro-scale physical details.[15]

These remarks on behalf of role functionalism challenge the considerations reviewed in the preceding section pointing to the conclusion that the conjunction of functionalism (in fact, role functionalism) and the multiple realizability of mental states would undermine the scientific usefulness of mental properties. The reader is urged to think about whether the remarks by Ross and Spurrett constitute an adequate rebuttal to our earlier considerations. One point the reader should notice is this: It is questionable whether, as Ross and Spurrett claim, our considerations in favor of realizer functionalism imply that we will get "bogged down in micro-scale physical details." Realizers are not necessarily, and not usually, individuated at the microphysical level.

Perhaps it might be argued that the actual practices and accomplishments of cognitive science and other special sciences go to show the emptiness of the essentially philosophical and a priori arguments of the preceding section. In spite of the heterogeneity of their underlying implementing mechanisms, functional role properties enter into laws and regularities that hold across diverse physical realizers. Ned Block, for example, has given some examples of psychological laws—in particular, those regarding stimulus generalization (due to the psychologist Roger Shepard)—that evidently seem to hold for all sorts of organisms and systems.[16] How these empirical results are to be correctly interpreted and understood, however, is an open question. The reader should keep in mind that an illusion of a systematic psychology and cognitive science may have been created by the fact that much of the research in these sciences focus on humans and related species. It is difficult to imagine a global scientific theory of, say, perception or memory as such, for all actual and nomologically possible psychological-cognitive systems, regardless of their modes of physical realization. A more detailed discussion of these issues takes us beyond core philosophy of mind and into the philosophy of psychology and cognitive science in a serious way. This is a good topic to reflect on for readers with an interest and background in these sciences.

For Further Reading

For statements of causal-theoretical functionalism, see David Lewis, "Psychophysical and Theoretical Identifications," and David Armstrong, "The Nature of Mind." Recommended also are Sydney Shoemaker, "Some Varieties of Functionalism," and Ned Block, "What Is Functionalism?"

Hilary Putnam, who was the first to articulate functionalism, has become one of its most severe critics; see his *Representation and Reality*, especially chapters 5 and 6. For other criticisms of functionalism, see Ned Block, "Troubles with Functionalism"; Christopher S. Hill, *Sensations: A Defense of Type Materialism*, chapter 3; and John R. Searle, *The Rediscovery of the Mind.* On the problem of qualia, see chapters 9 and 10 in this book and the suggested readings therein.

On the causal powers of functional properties, see Ned Block, "Can the Mind Change the World?"; Jaegwon Kim, *Mind in a Physical World*, chapter 2; Brian McLaughlin, "Is Role Functionalism Committed to Epiphenomenalism?"

The most influential statement of the multiple realization argument is Jerry Fodor, "Special Sciences, or the Disunity of Science as a Working Hypothesis." For the implications of multiple realization for cognitive-behavioral science, see Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction." For replies, see Ned Block, "Anti-Reductionism Slaps Back," and Jerry Fodor, "Special Sciences: Still Autonomous After All These Years." For a defense of cognitive science, see Don Ross and David Spurrett, "What to Say to a Skeptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists." For further discussion, see Gene Witmer, "Multiple Realizability and Psychological Laws: Evaluating Kim's Challenge."

Notes

1. This corresponds to machine functionalism's reference to the entire machine table of a Turing machine in characterizing its "internal" states. More below.

2. See David Lewis, "How to Define Theoretical Terms," and "Psychophysical and Theoretical Identifications."

3. Ramsey's original construction was in a more general setting of "theoretical" and "observational" terms rather than "psychological" and "physical-behavioral" terms. For details, see Lewis, "Psychophysical and Theoretical Identifications."

4. Here we follow Ned Block's method (rather than Lewis's) in his "What Is Functionalism?"

5. These remarks are generally in line with the "theory theory" of commonsense psychology. There is a competing account, the "simulation theory," according to which our use of commonsense psychology is not a matter of possessing a theory and applying its laws and generalizations but of "simulating" the psychology of others, using ourselves as models. See Robert M. Gordon, "Folk Psychology as Simulation," and Alvin I. Goldman, *Simulating Minds*. Prima facie, the simulation approach to folk psychology creates difficulties for the Ramsey-Lewis functionalization of mental terms. However, the precise implications of the theory need to be explored further.

6. The extension of a predicate, or concept, is the set of all things to which the predicate, or the concept, applies. So the extension of "human" is the set of all human beings. The extension of "unicorn" is the empty (or null) set.

7. For such a view, see Paul Churchland, "Eliminative Materialism and the Propositional Attitudes."

8. But it is difficult to imagine how the belief-desire-action principle *could* be shown to be empirically false. It has been argued that this principle is a priori true and hence resists empirical falsification. However, not all principles of vernacular psychology need to have the same status. It is possible that there is a core set of principles of vernacular psychology that can be considered a priori true and that suffices as a basis of the application of the Ramsey-Lewis method.

9. See Ned Block, "Troubles with Functionalism."

10. On the possibility of qualia inversion, see Sydney Shoemaker, "Inverted Spectrum"; Ned Block, "Are Absent Qualia Impossible?"; C. L. Hardin, *Color for Philosophers*; and Martine Nida-Rümelin, "Pseudo-Normal Vision: An Actual Case of Qualia Inversion?"

11. Being a carburetor is a functional property defined by a job description ("mixer of air and gasoline vapors" or some such), and a variety of physical devices can serve this purpose.

12. See, for example, Jerry Fodor, *Psychosemantics*, chapter 2.

13. For further discussion, see Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction." For replies, see Ned Block, "Anti-Reductionism Slaps Back," and Jerry Fodor, "Special Sciences: Still Autonomous After All These Years."

14. These terms are borrowed from Don Ross and David Spurrett, "What to Say to a Skeptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists." The discussion here is indebted to this article. The distinction between role and realizer functionalism closely parallels (is identical with?) Ned Block's distinction between the functional-state identity theory and the functional specification theory in his "What Is Functionalism?" Brian McLaughlin calls realizer functionalism "filler functionalism."

15. Don Ross and David Spurrett, "What to Say to a Skeptical Metaphysician."

16. Ned Block, "Anti-Reductionism Slaps Back."