Notes

1. To be more precise, the difference here must be understood to be a difference in the two states' "intrinsic" properties—for example, how they feel or look—not a difference in their "extrinsic" or "relational" properties, such as one of the states occurring south of Boston and the other to its north.

2. See René Descartes, *The Passions of the Soul.*

3. See Thomas H. Huxley, "On the Hypothesis That Animals Are Automata, and Its History."

4. Samuel Alexander, *Space, Time, and Deity.* Vol. 2, p. 47.

5. J. J. C. Smart, "Sensations and Brain Processes." U. T. Place's "Is Consciousness a Brain Process?," published in 1956, predates Smart's article as perhaps the first modern statement of the identity theory.

6. Smart, "Sensations and Brain Processes," p. 117 (in the reprint version in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil. Emphasis in original).

7. See the entry "William of Ockham" in the *Macmillan Encyclopedia of Philosophy*, 2nd ed.

8. Herbert Feigl, "The 'Mental' and the 'Physical,'" p. 428.

9. See Gilbert Harman, "The Inference to the Best Explanation." For a critique of the principle, see Bas Van Fraassen, *Laws and Symmetry.*

10. This example comes from Christopher S. Hill, *Sensations: A Defense of Type Materialism*, p. 24.

11. This is substantially the form in which Brian McLaughlin formulates his explanatory argument. See his "In Defense of New Wave Materialism: A Response to Horgan and Tienson." Hill (see note 10) and McLaughlin are two leading proponents of this form of the explanatory argument. See also Andrew Melnyck, *A Physicalist Manifesto.*

12. Ned Block and Robert Stalnaker, "Conceptual Analysis, Dualism, and the Explanatory Gap," p. 24.

13. This approach to events is usually associated with Donald Davidson; see Davidson, "The Individuation of Events."

14. A prime example of token physicalism is Donald Davidson's "anomalous monism." See Davidson, "Mental Events."

15. See Jaegwon Kim, "Events as Property Exemplifications."

16. This objection is worked out in detail in Jerome Shaffer, "Mental Events and the Brain."

17. See Smart, "Sensations and Brain Processes."

18. See, for example, Brian Loar, "Phenomenal States."

19. See Saul Kripke, *Naming and Necessity*, especially lecture 3, in which arguments against the identity theory are presented.

20. For further discussion of these issues, see the essays in *Conceivability and Possibility*, edited by Tamar Szabo Gendler and John Hawthorne.

21. The terms "variably realizable" and "variable realization" are commonly used by British writers.

5

Mind as a Computing Machine

Machine Functionalism

In 1967 Hilary Putnam published a paper of modest length titled "Psychological Predicates."[1] This paper changed the debate in philosophy of mind in a fundamental way by doing three remarkable things: First, it quickly brought about the decline and fall of type physicalism, in particular, the psychoneural identity theory. Second, it ushered in functionalism, which has since been a highly influential—arguably the preeminent—position on the nature of mind. And third, it helped to install antireductionism as the received view on the nature of mental properties and other "higher-level" properties of the special sciences. Psychoneural type physicalism, which had been promoted as the only view of mentality properly informed by the best contemporary science, turned out to be unexpectedly short-lived, and by the mid-1970s most philosophers had abandoned reductionist physicalism not only as a view about psychology but as a doctrine about all special sciences.[2]

All this was the work of a single idea: *the multiple realizability of mental properties.* We have already discussed it as an argument against the psychoneural identity theory and, more generally, as a difficulty for type physicalism (chapter 4). What sets the multiple realization argument apart from numerous other objections to the psychoneural identity theory is the fact that it gave birth to a new conception of the mental that has played a key

role in shaping a widely shared view of the nature and status of cognitive science and psychology.

Multiple Realizability and the Functional Conception of Mind

Perhaps not many of us now believe in angels—purely spiritual and immortal beings supposedly with a full mental life. Angels, as traditionally conceived, are wholly immaterial beings with knowledge and belief who can experience emotions and desires and are capable of performing actions. The idea of such a being may be a perfectly coherent one, like the idea of a unicorn or Bigfoot, but there is no evidence that there are beings fitting this description, just as there are no unicorns and probably no Bigfoot. So like unicorns but unlike married bachelors or four-sided triangles, there seems nothing conceptually impossible about angels. If the idea of an angel with beliefs, desires, and emotions is a consistent one, that would show that there is nothing in the idea of mentality as such that precludes purely nonphysical, wholly immaterial beings with psychological states.[3]

It seems, then, that we cannot set aside the possibility of immaterial realizations of mentality as a matter of an a priori conceptual fact.[4] Ruling out such a possibility requires commitment to a substantive metaphysical thesis, perhaps something like this:

> *Realization Physicalism.* If something *x* has some mental property M (or is in mental state M) at time *t*, then *x* is a material thing and *x* has M at *t* in virtue of the fact that *x* has at *t* some physical property P that realizes M in *x* at *t*.

It is useful to think of this principle as a way of characterizing physicalism.[5] It says that anything that exhibits mentality must be a physical system—for example, a biological organism. Although the idea of mentality permits nonphysical entities to instantiate mental properties, the world is so constituted, according to this thesis, that only physical systems, in particular, biological organisms, turn out to realize mental properties—perhaps because they are the only things that exist in space-time (see chapter 2). Moreover, the principle requires that every mental property be physically based; each occurrence of a mental property is due to the occurrence of a physical "realizer" of the mental property. A simple way of putting the point would be this: Minds, if they exist, must be embodied.

Notice that this principle provides for the possibility of multiple realization of mental properties. Mental property M—say, being in pain—may be such that in humans C-fiber activation realizes it but in other species (say, octopuses and reptiles) physiological mechanisms that realize pain may be vastly different. Perhaps there might be non-carbon-based or non-protein-based biological organisms with mentality, and we cannot a priori preclude the possibility that nonbiological electromechanical systems, like the "intelligent" robots and androids in science fiction, might be capable of having beliefs, desires, and even sensations. All this suggests an interesting feature of mental concepts: They seem to carry no constraint on the actual physical-biological mechanisms that, in a given system, realize or implement them. In this sense, psychological concepts are like concepts of artifacts. For example, the idea of an "engine" is silent on the actual physical mechanism that realizes it—whether it uses gasoline or electricity or steam and, if it is a gasoline engine, whether it is a piston or rotary engine, how many cylinders it has, whether it uses a carburetor or fuel injection, and so on. As long as a physical device is capable of performing a certain specified job—namely, that of transforming various forms of energy into mechanical force or motion—it counts as an engine. The concept of an engine is given by a *job description*, or *causal role*, not a description of mechanisms that execute the job. Many biological concepts are similar in the same respect: What makes an organ a heart is the fact that it pumps blood. The human heart may be physically very unlike hearts in, say, reptiles or birds, but they all count as hearts because of the job they do in the organisms in which they are found.

What, then, is the job description of pain? The capacity for experiencing pain under appropriate conditions—for example, when an organism suffers tissue damage—is critical to its chances for adaptation and survival. There are unfortunate people who congenitally lack the capacity to sense pain, and few of them survive into adulthood.[6] In the course of coping with the hazards presented by their environment, animal species must have had to develop pain mechanisms, what we may call "tissue-damage detectors," and it is plausible that different species, interacting with different environmental conditions and evolving independently of one another, have developed different mechanisms for this purpose. It is natural to expect to find diverse evolutionary solutions to the problem of developing a tissue-damage detector. As a start, then, we can think of pain as specified by the job description "tissue-damage detector"—a mechanism that is activated by tissue damage and whose activation in turn causes appropriate behavioral responses such as withdrawal, avoidance, and escape.

Thinking of the workings of the mind in analogy with the operations of a computing machine is commonplace, both in the popular press and in serious philosophy and cognitive science, and we will soon begin looking into the mind-computer analogy in detail. A computational view of mentality also shows that we must expect mental states to be multiply realized. We know that any computational process can be implemented in a variety of physically diverse computing machines. Not only are there innumerable kinds of electronic digital computers (in addition to the semiconductor-based machines we are familiar with, think of the vacuum-tube computers of olden days), but also computers can be built with wheels and gears (as in Charles Babbage's original "Analytical Engine") or even with hydraulically operated systems of pipes and valves, although these would be unacceptably slow (not to say economically prohibitive) by our current standards. And all of these physically diverse computers can be performing "the same computation," say, solving a given differential equation. If minds are like computers and mental processes—in particular, cognitive processes—are, at bottom, computational processes, we should expect no prior constraint on just how minds and mental processes are physically implemented. Just as vastly different physical devices can execute the same computational program, so vastly different biological or physical structures should be able to subserve the same psychological processes. This is the core of the functionalist conception of the mind.

What these considerations point to, according to some, is the *abstractness* or *formality* of psychological properties in relation to physical or biological properties: Psychological kinds abstract from the physical and biological details of organisms so that states that are vastly different from a physicochemical point of view can fall under the same psychological kind, and organisms and systems that are widely diverse biologically and physically can instantiate the same psychological regularities. To put it another way, psychological kinds seem to concern *formal* patterns or structures of events and processes rather than their material constitutions or implementing physical mechanisms.[7] Conversely, the same physical structure, depending on the way it is causally embedded in a larger system, can subserve different psychological capacities and functions (just as the same computer chip can be used for different computational functions in various subsystems of a computer). After all, most neurons, it has been argued, are pretty much alike and largely interchangeable.[8]

What is it, then, that binds together all the physically diverse instances of a given mental kind? What do all pains—pains in humans, pains in canines, pains in octopuses, and pains in Martians—have in common in virtue of which they all fall under a single psychological kind, pain?[9] That is, what is the *principle of individuation* for mental kinds?

Let us first see how the type physicalist and the behaviorist answer this question. The psychoneural type physicalist says this: What all pains have in common that makes them instances of pain is a certain neurobiological property, namely, being an instance of C-fiber excitation (or some such state). That is, for the type physicalist, a mental kind is a physical kind (a neurobiological kind, for the psychoneural identity theorist). You could guess how the behaviorist answers the question: What all pains have in common is a certain behavioral property—or to put it another way, two organisms are both in pain at a time just in case at that time they exhibit, or are disposed to exhibit, the behavior patterns definitive of pain (for example, escape behavior, withdrawal behavior, and so on). For the behaviorist, then, a mental kind is a behavioral kind.

If you take the multiple realizability of mental states seriously, you will reject both these answers and opt for a "functionalist" conception. The main idea is that what is common to instances of a mental state must be sought at a higher level of abstraction. According to functionalism, a mental kind is a *functional kind*, or a *causal-functional kind*, since the "function" involved is to fill a certain causal role.[10] Let us go back to pain as a tissue-damage detector.[11] The concept of a tissue-damage detector is a *functional concept*, a concept specified by a job description, as we said: Any device is a tissue-damage detector for an organism just in case it can reliably respond to occurrences of tissue damage in the organism and transmit this information to other subsystems so that appropriate behavioral responses are produced. Functional concepts are ubiquitous: What makes something a mouse trap, a carburetor, or a thermometer is its ability to perform a certain function, not any specific physicochemical structure or mechanism. These concepts are specified by the functions that are to be performed, not by structural blueprints. Many concepts, in ordinary discourse and in the sciences, seem to be functional concepts in this sense; even many chemical and biological concepts (for example, catalyst, gene, heart) appear to have an essentially functional component.

To return to pain as a tissue-damage detector: Ideally, every instance of tissue damage, and nothing else, should activate this mechanism—turn it on—and this must further trigger other mechanisms with which it is hooked up, leading finally to behavior that will in normal circumstances spatially separate the damaged part, or the whole organism, from the external cause of the

damage. Thus, the concept of pain is defined in terms of its function, and the function involved is to serve as a *causal intermediary* between typical pain inputs (tissue damage, trauma, and so on) and typical pain outputs (winces, groans, avoidance behavior, and so on). Moreover, functionalism makes two significant additions. First, among the causal conditions that activate the pain mechanism are other mental states (for example, you must be normally alert and not be absorbed in another activity, like intense competitive sports). Second, the outputs of the pain mechanism can include mental states as well (such as a sense of distress or a desire to be rid of the pain). The functionalist says that this is generally true of all mental kinds: Mental kinds are causal-functional kinds, and what all instances of a given mental kind have in common is the fact that they serve a certain *causal role* distinctive of that kind. As David Armstrong has put it, the concept of a mental state is that of an internal state apt to be caused by certain sensory inputs and apt to cause certain behavioral outputs.

Functional Properties and Their Realizers: Definitions

It will be useful to have explicit definitions of some of the terms we have been using more or less informally, relying on examples and intuitions. Let us begin with a formal characterization of a functional property:

> F is a *functional property* (or kind) just in case F can be characterized by a definition of the following form:
>
> For something *x* to have F (or to be an F) = $_{def}$ for *x* to have some property P such that C(P), where C(P) is a specification of the causal work that P is supposed to do in *x*.

We may call a definition having this form a "functional" definition. "C(P)," which specifies the causal role of F, is crucial. What makes a functional property the property it is, is the causal role associated with it; that is to say, F and G are the same functional property if and only if the causal role associated with F is the same as that associated with G. The term "causal work" in the above schema of functional definitions should be understood broadly to refer to "passive" as well as "active" work: For example, if tissue damage causes P to instantiate in an organism, that is part of P's causal work.

Thus, P's causal work refers to the *causal relations* involving the instances, or occurrences, of P in the organism or system in question.

Now we can define what it is for a property to "realize," or be a "realizer" of, a functional property:

> Let F be a functional property defined by a functional definition, as above. Property Q is said to *realize* F, or be a *realizer* or a *realization* of F, in system *x* if and only if C(Q), that is, Q fits the specification C in *x* (which is to say, Q in fact performs the specified causal work in system *x*).

Note that the definiens (the right-hand side) of a functional definition does not mention any particular property P that *x* has (when it has F); it only says that *x* has "some" property P fitting description C. In logical terminology, the definiens "quantifies over" properties (it in effect says, "There exists some property P such that *x* has P and C(P)."). For this reason, functional properties are called "second-order" properties, with the properties quantified over (that is, properties eligible as instances of P) counting as "first-order" properties; they are second-order properties of a special kind—namely, those that are defined in terms of causal roles.

Let us see how this formal apparatus works. Consider the property of being a mousetrap. It is a functional property because it can be given the following functional definition:

> *x* is a mousetrap = $_{def}$ *x* has some property P such that P enables *x* to trap and hold or kill mice.

The definition does not specify any specific P that *x* must have; the causal work to be done obviously can be done in many different ways. There are the familiar spring-loaded traps, and there are wire cages with a door that slams shut when a mouse enters; we can imagine high-tech traps with an optical sensor and all sorts of other devices. This means that there are many—in fact, indefinitely many—"realizers" of the property of being a mousetrap; that is, all sorts of physical mechanisms can be mousetraps.

Functionalism and Behaviorism

Both functionalism and behaviorism speak of sensory input and behavioral output—or "stimulus" and "response"—as central to the concept of

mentality. In this respect, functionalism is part of a broadly behavioral approach to mentality and can be considered a generalized and more sophisticated version of behaviorism. But there are also significant differences between them, of which the following two are the most important.

First, the functionalist takes mental states to be *real internal* states of an organism with causal powers; for an organism to be in pain is for it to be in an internal state (for example, a neurobiological state for humans) that is typically caused by tissue damage and that in turn typically causes winces, groans, and avoidance behavior. In contrast, the behaviorist eschews talk of internal states entirely, identifying mental states with actual or possible behavior. Thus, to be in pain, for the behaviorist, is to wince and groan or be disposed to wince and groan, but not, as the functionalist would have it, to be in some *internal state that causes* winces and groans.

Although both the behaviorist and the functionalist may refer to "behavioral dispositions" in speaking of mental states, what they mean by "disposition" can be quite different: The functionalist takes a "realist" approach to dispositions, whereas the behaviorist embraces an "instrumentalist" line. To see how realism and instrumentalism differ on this issue, consider how water solubility (that is, the disposition to dissolve in water) would be analyzed on each approach:

Instrumentalist analysis: x is soluble in water $=_{def}$ if x is immersed in water, x dissolves.

Realist analysis: x is soluble in water $=_{def}$ x is in a certain internal state S (that is, has a certain microstructure S) such that when x is immersed in water, S causes x to dissolve.

According to instrumentalism, therefore, the water solubility of a sugar cube is just the fact that a certain conditional ("if-then") statement holds for it; thus, on this view, water solubility is a "conditional" or "hypothetical" property of the sugar cube—that is, the property of *dissolving if immersed in water*. Realism, in contrast, takes solubility to be a categorical, presumably microstructural, internal state of the cube of sugar that is causally responsible for its dissolving when placed in water. (Further investigation might reveal the state to be that of having a certain crystalline molecular structure.) Neither analysis requires the sugar cube to be placed in water or actually to be dissolving in order to be water-soluble. However, we may note the following difference: If x dissolves in water and y does not, the realist will give a causal explanation of this difference in terms of a difference in their microstructure. For the instrumentalist, the difference may just be a brute fact: It is just that the conditional "if placed in water, it dissolves" holds true for x but not for y, a difference that need not be grounded in any further differences between x and y.

In speaking of mental states as behavioral dispositions, then, the functionalist takes them as actual inner states of persons and other organisms that in normal circumstances cause behavior of some specific type under certain specified input conditions. In contrast, the behaviorist takes them merely as input-output, or stimulus-response, correlations. Many behaviorists (especially methodological behaviorists) think that speaking of mental states as "inner causes" of behavior is scientifically unmotivated and philosophically unwarranted.[12]

The second significant difference between functionalism and behaviorism, one that gives the former a substantially greater theoretical power, is in the way "input" and "output" are construed for mental states. For the behaviorist, input and output consist entirely of observable physical stimulus conditions and observable behavioral responses. As briefly noted earlier, the functionalist allows reference to other *mental states* in the characterization of a given mental state. It is a crucial part of the functionalist conception of a mental state that its typical causes and effects can, and often do, include other mental states. Thus, for a ham sandwich to cause you to want to eat it, you must believe it to be a ham sandwich; a bad headache can cause you a feeling of distress and a desire to call your doctor.

The two points that have just been reviewed are related: If you think of mental states as actual inner states of psychological subjects, you would regard them as having real causal powers, powers to cause and be caused by other states and events, and there is no obvious reason to exclude mental states from figuring among the causes or effects of other mental states. In conceiving mentality this way, the functionalist is espousing *mental realism*—a position that considers mental states as having a genuine ontological status and counts them among the phenomena of the world with a place in its causal structure. Mental states are real for the behaviorist too, but only as behaviors or behavioral dispositions; for him, there is nothing mental over and above actual and possible behavior. For the functionalist, mental states are inner causes of behavior, and as such they are "over and above" behavior.

Including other mental events among the causes and effects of a given mental state is part of the functionalist's general conception of mental states as forming a complex causal network anchored to the external world at various points. At these points of contact, a psychological subject interacts with

the outside world, receiving inputs and emitting outputs. And the identity of a given mental kind, whether it is a sensation like pain or a belief that it is going to rain or a desire for a ham sandwich, depends solely on the place it occupies in the causal network. That is, what makes a mental event the kind of mental event it is, is the way it is causally linked to other mental-event kinds and input-output conditions. Since each of these other mental-event kinds in turn has its identity determined by *its* causal relations to other mental events and to inputs and outputs, the identity of each mental kind depends ultimately on the whole system—its internal structure and the way it is causally linked to the external world via sensory inputs and behavior outputs. In this sense, functionalism gives us a *holistic* conception of mentality.

This holistic approach enables functionalism to sidestep one of the principal objections to behaviorism. This is the difficulty we saw earlier: A desire issues in overt behavior only when combined with an appropriate belief, and similarly, a belief leads to appropriate behavior only when a matching desire is present. For example, a person with a desire to eat an apple will eat an apple that is presented to her only if she believes it to be an apple (she would not eat it if she thought it was a fake wooden apple); a person who believes that it is going to rain will take an umbrella only if she has a desire not to get wet. As we saw, this apparently makes it impossible to define desire without reference to belief or define belief without reference to desire. The functionalist would say that this only points to the holistic character of mental states: It is an essential feature of a desire that it is the kind of internal state that in concert with an appropriate belief causes a certain behavior output, and similarly for belief and other mental states.

But doesn't this make the definitions circular? If the concept of desire cannot be defined without reference to belief, and the concept of belief in turn cannot be explained without reference to desire, how can either be understood at all? We will see below (chapter 6) how the holistic approach of functionalism deals with this question.

Turing Machines

Functionalism was originally formulated by Putnam in terms of "Turing machines," mathematically characterized computing machines due to the British mathematician-logician Alan M. Turing.[13] Although it is now customary to formulate functionalism in terms of causal-functional roles—as we have done and will do in more detail in the next chapter—it is instructive to begin our systematic treatment of functionalism by examining the Turing-machine version of functionalism, usually called machine functionalism. This also gives us the background needed to explore the idea that the workings of the mind are best understood in terms of the operations of a computing machine—that is, the computational view of the mind (or computationalism for short).

A Turing machine is made up of four components:

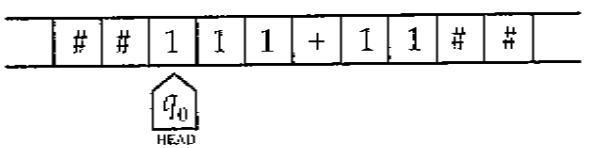
1. A *tape* divided into "squares" and unbounded in both directions
2. A *scanner-printer* ("head") positioned at one of the squares of the tape at any given time
3. A finite set of *internal states* (or *configurations*), $q_0, \ldots, q_n$
4. A finite *alphabet* consisting of symbols, $b_1, \ldots, b_m$

One and only one symbol appears on each square. (We may think of the blank as one of the symbols.)

The general operations of the machine are as follows:

A. At each time, the machine is in one of its internal states, q_i, and its head is scanning a particular square on the tape.
B. What the machine does at a given time t is completely determined by its internal state at t and the symbol its head is scanning at t.
C. Depending on its internal state and the symbol being scanned, the machine does three things:
   (1) Its head replaces the symbol with another (possibly the same) symbol of the alphabet. (To put it another way, the head erases the symbol being scanned and prints a new one, which may be the same as the erased one.)
   (2) Its head moves one square to the right or to the left (or halt, with the computation completed).
   (3) The machine enters into a new internal state.

Let us consider a Turing machine that adds positive integers in the unary notation. (In this notation, number n is represented as a sequence of n strokes, each stroke occupying one square.) Consider the following picture in which the problem of adding 3 and 2 is presented to the machine, which is to be started off with its head in state q_0 and scanning the first digit:



(The "scratch" symbol, #, marks the boundaries of the problem.) We want to "program" this Turing machine in such a way that when the computation is completed, the machine halts with a sequence of five consecutive strokes showing on the tape, like this:

	#	#	1	1	1	1	1	#	#	#

It is easy to see that there are various procedures by which the machine could accomplish this. One simple way is to have the machine (or its head) move to the right looking for the symbol +, replace it with a stroke, keep moving right until it finds the right-most stroke, and when it does, erase it (that is, replace it with the scratch symbol #) and then halt. The following simple "machine table" is a complete set of instructions that defines our adder (call it TM_1):

	q_0	q_1
1	$1Rq_0$	#Halt
+	$1Rq_0$	
#	$\#Lq_1$	

Here is how we read this table. On the left-most column you find the symbols of the machine alphabet listed vertically, and the top row lists the machine's internal states. Each entry in the interior matrix is an *instruction*: It tells the machine what to do when it is scanning the symbol shown in the left-most column of that row and is in the internal state listed at the top of the column. For example, the entry $1Rq_0$, at the intersection of q_0 and 1, tells the machine: "If you are scanning the symbol 1 and are in internal state q_0, replace 1 with 1 (that is, leave it unchanged), move to the right by one square, and go into internal state q_0 (that is, stay in the same state)." The entry immediately below, $1Rq_0$, tells the machine: "If you are in state q_0 and scanning the symbol +, replace + with 1, move to the right by one square, and go into state q_0." The L in the bottom entry, $\#Lq_1$, means "move left by one square"; the entry in the right-most column, #Halt, means "If you are scanning 1 and in state q_1, replace 1 with # and halt." It is easy to see (the reader is asked to figure this out on her own) the exact sequence of steps our Turing machine will follow to compute the sum 3 + 2.

The machine table of a Turing machine is a complete and exhaustive specification of the machine's operations. We may therefore identify a Turing machine with its machine table. Since a machine table is nothing but a set of

instructions, this means that a Turing machine can be identified with a set of such instructions.

What sort of things are the "internal states" of a Turing machine? We talk about this general question later, but with our machine TM_1, it can be helpful to think of the specific machine states in the following intuitive way: q_0 is a + and # searching state—it is a state such that when TM_1 is in it, it keeps going right, looking for + and #, ignoring any 1s it encounters. Moreover, if the machine is in q_0 and finds a +, it replaces it with a 1 and keeps moving to the right, while staying in the same state; when it scans a # (thereby recognizing the right-most boundary of the given problem), it backs up to the left and goes into a new state q_1, the "print # over 1 and then halt" state. When TM_1 is in this state, it will replace any 1 it scans with a # and halt. Thus, each state "disposes" the machine to do a set of specific things depending on the symbol being scanned (which therefore can be likened to sensory input).

But this is not the only Turing machine that can add numbers in unary notation; there is another one that is simpler and works faster. It is clear that to add unary numbers it is not necessary for the machine to determine the right-most boundary of the given problem; all it needs to do is to erase the initial 1 being scanned when it is started off, and then move to the right to look for + and replace it with a 1. This is TM_2, with the following machine table:

	q_0	q_1
1	$\#Rq_1$	$1Rq_1$
+		1Halt
#		

We can readily build a third Turing machine, TM_3, that will do subtractions in the unary notation. Suppose the following subtraction problem is presented to the machine:

#	#	b	1	1	1	1	−	1	1	b	#	#

(Symbol b is used to mark the boundaries of the problem.) Starting the machine in state q_0 scanning the initial 1, we can write a machine table that computes $n - m$ by operating like this:

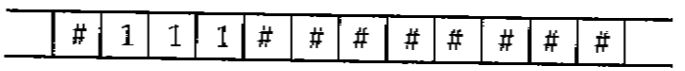
1. The machine starts off scanning the first 1 of n. It goes to the right until it locates m, the number being subtracted. (How does it recognize it has located m?) It then erases the first 1 of this number

(replacing it with a #), goes left, and erases the last 1 of n (again replacing it with a #).

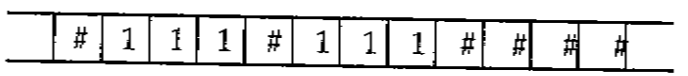
2. The machine then goes right and repeats step 1 again and again, until it exhausts all the 1s in m. (How does the machine "know" that it has done this?) We then have the machine move right until it locates the subtraction sign –, which it erases (that is, replaces it with a #), and then halt. (If you like tidy output tapes, you may have the machine erase the bs before halting.)
3. If the machine runs out of the first set of strokes before it exhausts the second set (this means that $n < m$), we can have the machine print a certain symbol, say ?, to mean that the given problem is not well-defined. We must also provide for the case where $n = m$.

The reader is invited to write out a machine table that implements these operations.

We can also think of a "transcription machine," TM_4, that transcribes a given string of 1s to its right (or left). That is, if TM_4 is presented with the following tape to begin its computation,



it ends with the following configuration of symbols on its tape:



The interest of the transcription machine lies in how it can be used to construct a multiplication machine, TM_5. The basic idea is simple: We can get $n \times m$ by transcribing the string of n 1s m times (that is, transcribing n repeatedly using m as a counter). The reader is encouraged to write a machine table for TM_5.

Since any arithmetical operation (squaring, taking the factorial, and so on) on natural numbers can be defined in terms of addition and multiplication, it follows that there is a Turing machine that computes any arithmetical operation. In fact, it can be shown that any computation performed by any computer can be done by a Turing machine. That is, being computable and being computable by a Turing machine turn out to be equivalent notions.

We can think of a Turing machine with two separate tapes (one for input, on which the problem to be computed is presented, and the other for actual computation and the final output) and two separate heads (one for scanning

and one for printing). This helps us to think of a Turing machine as receiving "sensory stimuli" (the symbols on the input tape) through its scanner ("sense organ") and emitting specific behaviors in response (the symbols printed on the output tape by its printer head). It can be shown that any computation that can be done by a two-tape machine or a machine with any finite number of tapes can be done by a one-tape machine. So adding more tapes does not strengthen the computing power of Turing machines or substantively enrich the concept of a Turing machine.

Turing also showed how to build a "universal machine," which is like a general-purpose computer in that it is not dedicated to the computation of a specific function but can be programmed to compute any function you want. On the input tape of this machine, you specify two things: the machine table of the desired function in some standard notation that can be read by the universal machine and the values for which the function is to be computed. The universal machine is programmed to read any machine table and carry out the computation in accordance with the instructions of the machine table.

The notion of a Turing machine can be generalized to yield the notion of a *probabilistic automaton*. As you recall, each instruction of a Turing machine is *deterministic*: Given the internal state and the symbol being scanned, the immediate next operation is wholly and uniquely determined. An instruction of a probabilistic, or stochastic, automaton has the following general form: Given internal state q_i and scanned symbol b_j:

1. Print b_k with probability r_1, or print b_l with probability r_2, . . . , or print b_m with probability r_n (where the probabilities add up to 1).
2. Move R with probability r_1, or move L with probability r_2 (where the probabilities add up to 1).
3. Go into internal state q_j with probability r_1, or into q_k with probability r_2, . . . , or into q_m with probability r_n (again, the probabilities adding up to 1).

A machine can be made probabilistic in one or more of these three dimensions. The operations of a probabilistic automaton, therefore, are not deterministic; the current internal state of the machine and the symbol it is scanning do not together uniquely determine what the machine will do next. However, the behavior of such a machine is not random or arbitrary either: There are fixed and stable probabilities describing the machine's operations. If we are thinking of a machine that describes the behavior of an actual

psychological subject, a probabilistic machine may be more realistic than a deterministic one; however, we may note the fact that it is generally possible to construct a deterministic machine that simulates the behavior of a probabilistic machine to any desired degree of accuracy, which makes probabilistic machines theoretically dispensable.

Physical Realizers of Turing Machines

Suppose that we give the machine table for our simple adding machine, TM_1, to an engineering class as an assignment: Each student is to build an actual physical device that will do the computations as specified by its machine table. What we are asking the students to build, therefore, are "physical realizers" of TM_1—real-life physical computing machines that will operate in accordance with the machine table of TM_1. We can safely predict that a huge, heterogeneous variety of machines will be turned in. Some of them may really look and work like the Turing machine as described: They will have a paper tape neatly divided into squares, with an actual physical "head" that can read, erase, and print symbols. Some will perhaps use magnetic tapes and heads that read, write, and erase electrically. Some machines will have no "tapes" or "heads" but instead use spaces on a computer disk or memory locations in its CPU to do the computation. A clever student with a sense of humor (and lots of time and other resources) might try to build a hydraulically operated device with pipes and valves instead of wires and switches. The possibilities are endless.

But what exactly is a physical realizer of a Turing machine? What makes a physical device a *realizer of* a given Turing machine? First, the symbols of the machine's alphabet must be given concrete physical embodiments; they could be blotches of ink on paper, patterns of magnetized iron particles on plastic tape, electric charges in capacitors, or what have you. Whatever they are, the physical device that does the "scanning" must be able to "read" them—that is, differentially respond to them—with a high degree of reliability. This means that the physical properties of the symbols place a set of constraints on the physical design of the scanner, but these constraints need not, and usually will not, determine a unique design; a great multitude of physical devices are likely to be adequate to serve as a scanner for the given set of physically embodied symbols. The same considerations apply to the machine's printer and outputs as well: The symbols the machine prints on its output tape (we are thinking of a two-tape machine) must be given physical

shapes, and the printer must be designed to produce them on demand. The printer, of course, does not have to "print" anything in a literal sense; the operation could be wholly electronic, or the printer could be a speaker that vocalizes the output or an LCD monitor that visually displays it (then saves it for future computational purposes).

What about the "internal states" of the machine? How are they physically realized? Consider a particular instruction on the machine table of TM_1: If the machine is in state q_0 and scanning a +, replace that + with a 1, move right, and go into state q_1. Assume that Q_0 and Q_1 are the physical states realizing q_0 and q_1, respectively. Q_0 and Q_1, then, must satisfy the following condition: An occurrence of Q_0, together with the physical scanning of +, must *physically cause* three physical events: (1) The physical symbol + is replaced with the physical symbol 1; (2) the physical scanner-printer (head) moves one square to the right (on the physical tape) and scans it; and (3) the machine enters state Q_1. In general, then, what needs to be done is to *replace the functional or computational relations* among the various abstract parameters (symbols, states, and motions of the head) mentioned in the machine table *with appropriate causal relations among the physical embodiments* of these parameters.

From the logical point of view, the internal states are only "implicitly defined" in terms of their relations to other parameters: q_j is a state such that if the machine is in it and scanning symbol b_k, the machine replaces b_k with b_i, moves R (that is, to the right), and goes into state q_h; if the machine is scanning b_m, it does such and such; and so on. So q_j can be thought of as a function that maps symbols of the alphabet to the triples of the form $<b_i$, R (or L), $q_h>$. From the physical standpoint, Q_j, which realizes q_j, can be thought of as a *causal* intermediary between the physically realized symbols and the physical realizers of the triples—or equivalently, as a *disposition* to emit appropriate physical outputs (the triples) in response to different physical stimuli (the physical symbols scanned). This means that the intrinsic physical natures of the Qs that realize the qs are of no interest to us as long as they have the right causal properties; their intrinsic properties do not matter—or more accurately, they matter only to the extent that they affect the desired causal powers of the states and objects that have them. As long as these states perform their assigned causal work, they can be anything you please. Clearly, whether or not the Qs realize the qs depends crucially on how the tape, symbols, and so on, are physically realized; in fact, these are interdependent questions. It is plausible to suppose that, with some engineering ingenuity,

a machine could be rewired so that physical states realizing distinct machine states could be interchanged without affecting the operation of the machine.

We see, then, a convergence of two ideas: the functionalist conception of a mental state as a state occupying a certain specific causal role and the idea of a physical state realizing an internal state of a Turing machine. Just as, on the functionalist view, what makes a given mental state the kind of mental state it is is its causal role with respect to sensory inputs, behavior outputs, and other mental states, so what makes a physical state the realizer of a given machine state is its causal relations to inputs, outputs, and other physical realizers of the machine's internal states. This is why it is natural for functionalists to look to Turing machines for a model of the mind.

Let S be a physical system (which may be an electromechanical device like a computer, a biological organism, a business organization, or anything else), and assume that we have adopted a vocabulary to describe its inputs and outputs. That is, we have a specification of what is to count as the inputs it receives from its surroundings and what is to count as its behavioral outputs. Assume, moreover, that we have specified what states of S are to count as its "internal states." We will say that a Turing machine M is a *machine description* of system S, relative to a given input-output specification and a specification of the internal states, just in case S realizes M relative to the input-output and internal state specifications. Thus, the relation of *being a machine description of* is the converse of the relation of *being a realizer (or realization) of*. We can also define a concept that is weaker than machine description: Let us say that a Turing machine M is a *behavioral description* of S (relative to an input-output specification) just in case M provides a correct description of S's input-output correlations. Thus, every machine description of S is also a behavioral description of S, but the converse does not in general hold. M can give a true description of the input-output correlations characterizing S, but its machine states may not be realized in S, and S's inner workings (that is, its computational processes) may not correctly mirror the functional-computational relationships given by M's machine table. In fact, there may be another Turing machine M*, distinct from M, that gives a correct machine description of S. It follows, then, that *two physical systems that are input-output equivalent may not be realizations of the same Turing machine*. (The pair of adding machines TM_1 and TM_2 illustrates such a situation.)

Machine Functionalism: Motivations and Claims

Machine functionalists claim that we can think of the mind as a Turing machine (or a probabilistic automaton). This of course needs to be filled out, but from the preceding discussion it should be pretty clear how the story will go. The central idea is that what it is for something to have mentality—that is, to have a psychology—is for it to be a physically realized Turing machine of appropriate complexity, with its mental states (that is, mental-state types) identified with the internal states of the machine table. Another way of explaining this idea is to use the notion of machine description: An organism has mentality just in case there is a Turing machine of appropriate complexity that is a machine description of it, and its mental-state kinds are to be identified with the internal states of that Turing machine. All this is, of course, relative to an appropriately chosen input-output specification, since you must know, or decide, what is to count as the organism's inputs and outputs before you can determine what Turing machine (or machines) it can be said to instantiate.

Let us consider the idea that *the psychology of an organism* can be represented by a Turing machine, an idea that is commonly held by machine functionalists.[14] Let V be a complete specification of all possible inputs and outputs of a psychological subject S, and let C be all actual and possible input-output correlations of S (that is, C is a complete specification of which input applied to S elicits which output, for all inputs and outputs listed in V). In constructing a *psychology* for S, we are trying to formulate a *theory* that gives a perspicuous systematization of C by positing a set of internal states in S. Such a theory *predicts* for any input applied to S what output will be emitted by S and also *explains* why that particular input will elicit that particular output. It is reasonable to suppose that for any behavioral system complex enough to have a psychology, this kind of systematization is not possible unless we advert to its internal states, for we must expect that the same input applied to S does not always prompt S to produce the same output. The actual output elicited by a given input depends, we must suppose, on the internal state of S at that time.

Before we proceed further, it is necessary to modify our notion of a Turing machine in one respect: The internal states, *q*s, of a Turing machine are *total* states of the machine at a given time, and the Qs that are their physical realizers are also *total* physical states at a time of the physically realized machine.

This means that the Turing machines we are talking about are not going to look very much like the psychological theories we are familiar with; the states posited by these theories are seldom, if ever, total states of a subject at a time. But this is a technical problem, something that we assume can be remedied with a finer-grained notion of an "internal state." We can then think of a total internal state as made up of these "partial" states, which combine in different ways to yield different total states. This modification should not change anything essential in the original conception of a Turing machine. In the discussion to follow, we use this modified notion of an internal state in most contexts.

To return to the question of representing the psychology of a subject S in terms of a Turing machine: What Turing machine, or machines, is adequate as a description of S's psychology? Evidently, any adequate Turing machine must be a behavioral description of S, in the sense defined earlier; that is, it must give a correct description of S's input-output correlations (relative to V). But as we have seen, there is bound to be more than one Turing machine—in fact, if there is one, there will be indefinitely more—that gives a correct representation of S's input-output correlations.

Since each of these machines is a correct behavioral description of our psychological subject S, they are all equally good as a *predictive instrument*. Although some of them may be easier to manipulate and computationally more efficient than others, they all predict the same behavior output for the same input conditions. This is a simple consequence of the notion of "behavioral description." In what sense, then, are they different Turing machines, and why do the differences matter?

It should be clear how behaviorally equivalent Turing machines, say, M_1 and M_2, can differ from each other. To say that they are different Turing machines is to say that their machine tables are different—that is how Turing machines are individuated. This means that when they are given the same input, M_1 and M_2 are likely to go through *different computational processes* to arrive at the same output. Each machine has a set of internal states—let us say $<q_0, q_1, \ldots, q_n>$ for M_1 and $<r_0, r_1, \ldots, r_m>$ for M_2. Let us suppose further that M_1 is a machine description of our psychological subject S, but M_2 is not. That is, S is a physical realizer of M_1 but not of M_2. This means that the computational relations represented in M_1, but not those represented in M_2, are mirrored in a set of causal relations among the physical-psychological states of S. So there are real physical (perhaps neurobiological) states in S, $<Q_0, Q_1, \ldots, Q_n>$, corresponding to M_1's internal states $<q_0, q_1, \ldots, q_n>$, and these Qs are causally hooked up to each other and to the physical scanner (sense organs) and the physical printer (motor mechanisms) in a way that ensures that for all computational processes generated by M_1, isomorphic causal processes occur in S. As we may say, S is a "causal isomorph" of M_1.

There is, then, a clear sense in which M_1 is, but M_2 is not, "psychologically real" for S, even though they are both accurate predictive theories of S's observable input-output behaviors. M_1 gives "the true psychology" of S in that, as we saw, S has a physical structure whose states constitute a causal system that mirrors the computational structure represented by the machine table of M_1 and in that the physical-causal operations of S form an isomorphic image of the computational operations of M_1. This makes a crucial difference when what we want is an *explanatory* theory, a theory that *explains why S does what it does under the given input conditions*. Suppose we say: When input *i* was applied to S, S emitted behavioral output *o* *because* it was in internal state Q. This can count as an explanation, it seems, only if the state appealed to—namely, Q—is a "real" state of the system. In particular, it can count as a *causal* explanation only if the state Q is what, in conjunction with *i*, caused *o*, and this cannot happen unless Q is actually a state of S. Now, it is clear that we can impute reality to Q only if it realizes an internal state of M_1, a Turing machine realized by S. In contrast, Turing machine M_2, which is not realized by S, has no "inner" psychological reality for S, even though it correctly captures all of S's input-output correlations. Although, like M_1, M_2 correlates input *i* with output *o*, the computational process whereby the correlation is effected does not reflect actual causal processes in S that lead from *i* to *o* (or physical embodiments thereof). The explanatory force of "S emitted *o* when it received input *i* because it was in state Q" derives from the causal relations involving Q and the physical embodiments of *o* and *i*.

The philosophical issues here depend, partly but critically, on the metaphysics of scientific theories you accept. If you think of scientific theories in general, or theories over some specific domain, merely as predictive instruments that enable us to infer or calculate further observations from the given data, you will not attach any existential significance to the posits of these theories—like the unobservable microparticles of theoretical physics and their (often quite strange) properties—and may regard them only as calculational aids in deriving predictions. A position like this is called "instrumentalism," or "antirealism," about scientific theory.[15] On such a view, the issue of "truth" does not arise for the theoretical principles, nor does the issue of "reality" for the entities and properties posited; the only thing that matters is the "empirical adequacy" of the theory—how accurately the theory works as a predictive device and how comprehensive its coverage is. If you accept an

instrumentalist stance toward psychological theory, therefore, any Turing machine that is a behavioral description of a psychological subject is good enough, exactly as good as any other behaviorally adequate description of it, although you may prefer some over others on account of manipulative ease and computational cost. If this is your view of the nature of psychology, you will dismiss as meaningless the question which of the many behaviorally adequate psychologies is "really true" of the subject.

But if you adopt the perspective of "realism" on scientific theories, or at any rate about psychology, you will not think all behaviorally adequate descriptions are psychologically adequate. An adequate psychology for the realist must have "psychological reality": That is, the internal states it posits must be the real states of the organism with an active role as causal intermediaries between sensory inputs and behavior outputs, and this means that only a Turing machine that is a correct machine description of the organism is an acceptable psychological theory of it. The simplest and most elegant behavioral description may not be the one that correctly describes the inner processes that cause the subject's observable behavior; there is no a priori reason to suppose that our subject is put together according to the specifications of the simplest and most elegant theory (whatever your standards of simplicity and elegance might be).

Why should one want to go beyond the instrumentalist position and insist on psychological reality? There are two related reasons: (1) Psychological states, namely, the internal states of the psychological subject posited by a psychology, must be regarded as real, as we saw, if we expect the theory to generate explanations, especially causal explanations, of behavior. And this seems to be the attitude of working psychologists: It is their common, almost universal, practice to attribute to their subjects internal states, capacities, functions, and mechanisms (for example, information storage and retrieval, mental imagery, preference structure) and to refer to them in formulating what they regard as causal explanations of overt behavior. Further, (2) it seems natural to expect—this seems true of most psychologists and cognitive scientists—to find actual neural-biological mechanisms that underlie the psychological states, capacities, and functions posited by correct psychological theories. Research in the neural sciences has had impressive successes—and we expect this to continue—in identifying physiological mechanisms that implement psychological capacities and functions. It is a reflection of our realistic stance toward psychological theorizing that we generally expect, and perhaps insist on, physiological foundations for psychological theories. The requirement that the correct psychology of an organism be a machine

description of it,[16] not merely a behaviorally adequate one, can be seen as an expression of a commitment to realism about psychological theory.

If the psychology of any organism can be represented as a Turing machine, it is natural to consider the possibility of using representability by a Turing machine to explicate, or define, what it is for something to have a psychology. As we saw, that precisely is what machine functionalism proposes: What it is for an organism, or system, to have a psychology—that is, what it is for an organism to have mentality—is for it to realize an appropriate Turing machine. It is not merely that anything with mentality has an appropriate machine description; machine functionalism makes the stronger claim that its having a machine description of an appropriate kind is *constitutive of* its mentality. This is a philosophical thesis about the nature of mentality: Mentality, or having a mind, consists in realizing an appropriate Turing machine. What makes us creatures with mentality, therefore, is the fact that we are Turing machines. Functionalism acknowledges that having a brain of a certain structural complexity is important to mentality, but the importance of the brain lies exactly in its being a physical Turing machine. It is our brain's computational powers, not its biological properties, that constitute our mentality. In short, our brain is our mind because it is a computing machine, not because it is composed of the kind of protein-based biological stuff it is composed of.

Machine Functionalism: Further Issues

Suppose that two systems, S_1 and S_2, are *in the same mental state* (at the same time or different times). What does this mean on the machine-functionalist conception of a mental kind? A mental kind, as you will remember, is supposed to be an internal state of a Turing machine (of an "appropriate kind"); so for S_1 and S_2 to be in the same state, there must be some Turing machine state q such that S_1 is in q and S_2 is also in q. But what does this mean?

S_1 and S_2 are both physical systems, and we know that they could be systems of very different sorts (recall multiple realizability). As physical systems, they have physical states (that is, they instantiate certain physical properties); to say that they are both in machine state q at time t is to say this: There are physical states Q_1 and Q_2 such that Q_1 realizes q in S_1, and Q_2 realizes q in S_2, and, at t, S_1 is in Q_1 and S_2 in Q_2. Multiple realizability tells us that Q_1 and Q_2 probably have not much in common qua physical states; one could be a biological state and the other an electromagnetic one. What binds the two states together is only the fact that in their respective systems they

implement the same internal machine state. That is to say, the two states play the same computational role in their respective systems.

But talk of "the same internal machine state q" makes sense only in relation to a given machine table. That is to say, internal states of a Turing machine are identifiable only relative to a given machine table: In terms of the layout of machine tables we used earlier, an internal state q is wholly characterized by the vertical column of instructions appearing under it. But these instructions refer to other internal states, say, q_i, q_j, and q_k, and if you look up the instructions falling under these, you are likely to find references back to state q. So these states are interdefined. What all this means is that *the sameness or difference of an internal state across different machine tables—that is, across different Turing machines—has no meaning.* It makes no sense to say of an internal state q_i of one Turing machine and a state q_k of another that q_i is, or that it is not, the same state as q_k; nor does it make sense to say of a physical state Q_j of a physically realized Turing machine that it realizes, or does not realize, the same internal machine state q as does a physical state Q_k of another physical machine, *unless the two physical machines are realizations of the same Turing machine.*

Evidently, then, the machine-functionalist conception of mental kinds has the following consequence: For any two subjects to be in the same mental state, they must realize the same Turing machine. But if they realize the same Turing machine, their total psychology must be identical. That is, on machine functionalism, two subjects' total psychology must be identical if they are to share even a single psychological state—or even to give meaning to the talk of their being, or not being, in the same psychological state. This sounds absurd: It does not seem reasonable to require that for two persons to share a mental state—say, the belief that snow is white—the total set of psychological regularities governing their behavior must be exactly identical. Before we discuss this issue further, we must attend to another matter, and this is the problem of how the inputs and outputs of a system are to be specified.

Suppose that two systems, S_1 and S_2, realize the same Turing machine; that is, the same Turing machine gives a correct machine description for each. We know that realization is relative to a particular input-output specification; that is, we must know what is to count as input conditions and what is to count as behavior outputs of the system before we can tell whether it realizes a given Turing machine. Let V_1 and V_2 be the input-output specifications for S_1 and S_2, respectively, relative to which they share the same machine description. Since the same machine table is involved, V_1 and V_2 must be isomorphic: The elements of V_1 can be correlated, one to one, with the elements of V_2 in a way that preserves their roles in the machine table. And we are assuming, of course, that our Turing machine has the appropriate complexity to qualify as a psychological system.

But suppose that S_1 is a real psychological system, perhaps a human (call him Larry), whereas S_2 is a computer, an electromechanical device (call it MAX). So the inputs and outputs specified by V_2 are the usual inputs and outputs appropriate for a computing machine, perhaps strings of symbols entered on the keyboard and strings of symbols on the monitor or its printout. Now, whether MAX can be considered a psychological system at all is a question we take up later, but granting it the full psychological status that we grant Larry should strike us as in effect *conflating a psychological subject with a computer simulation of it.* It is to refuse to acknowledge a distinction between a real thing and a computer simulation of a real thing. No one is likely to confuse the operation of a jet engine or the spread of rabies in wildlife with their computer simulations. It is difficult to believe that this distinction suddenly vanishes when we perform a computer simulation of the psychology of a person.

One thing that obviously seems wrong about our computer, MAX, as a psychological system when we compare it with Larry is its inputs and outputs: Although its input-output specification is isomorphic to Larry's, it seems entirely inappropriate for psychology. It may not be easy to characterize the differences precisely, but we would not consider inputs and outputs consisting merely of strings of symbols as appropriate for something with true mentality. Grinding out strings of symbols is not like the full-blown behavior that we see in Larry. For one thing, MAX's outputs have nothing to do with its survival or continued proper functioning, and its inputs do not have the function of providing MAX with information about its surroundings. As a result, MAX's outputs lack what may be called "teleological aptness" as a response to its inputs. All this makes it difficult to think of MAX's outputs as constituting real behavior or action, something that is necessary if we are to regard it as a genuine psychological system.

Qua realizations of a Turing machine, MAX and Larry are symmetrically related. If, however, we see here an asymmetry in point of mentality, it is clear that the nature of inputs and outputs is an important factor, and our considerations seem to show that for a system realizing a Turing machine to count as a psychological system, its input-output specification (relative to which it realizes the machine) must be *psychologically appropriate.* Exactly what this appropriateness consists in is an interesting and complex question that requires further exploration. In any case, the machine functionalist must

confront this question: Is it possible to give a characterization of this input-output appropriateness that is consistent with functionalism—in particular, without using mentalistic terms or concepts? Recall a similar point we discussed in connection with behaviorism: Not to beg the question, the behavior that the behaviorist is allowed to talk about in giving behavioristic definitions of mental concepts must be "physical behavior," not intentional action with an explicit or implicit mental component (such as reading the morning paper, being impolite to a waiter, or going to a concert). If your project is to get mentality out of behavior, your notion of behavior must not presuppose mentality.

The same restriction applies to the machine functionalist: Her project is to define mentality in terms of Turing machines and input-output relations. The additional tool she can make use of, something not available to the behaviorist, is the concept of a Turing machine with its "internal" states, but her input and output are subject to the same constraint—her input-output, like the behaviorist's, must be physical input-output. If this is right, it seems no easy task for the machine functionalist to distinguish, in a principled way, Larry's inputs-outputs from MAX's, and hence genuine psychological systems from their simulations. We pointed out earlier that Larry's outputs, given his inputs, seem *teleologically apt*, whereas MAX's do not. They have something to do with his proper functioning in his environment—coping with environmental conditions and changes and satisfying his needs and desires. But can this notion of teleology—purposiveness or goal-directedness—be explained in a psychologically neutral way, without begging the question? Perhaps, some biological-evolutionary story could be attempted, but it remains an open question whether such a bioteleological attempt will succeed. In any case, that would take us beyond machine functionalism proper. These considerations give credence to the idea that in order to have genuine mentality, a system must be embedded in a natural environment (ideally including other systems like it), interacting and coping with it and behaving appropriately in response to the ever-changing stimulus conditions it encounters.

Let us now return to the question of whether machine functionalism is committed to the consequence that two psychological subjects can be in the same psychological state only if they have an identical total psychology. As we saw, the implication appears to follow from the fact that, on machine functionalism, being in the same psychological state is being in the same internal machine state and that the sameness of a machine state makes sense only in relation to the same Turing machine. What is perhaps worse, it seems to follow that it makes no sense to say that two psychological subjects are *not*

in the same psychological state unless they have an identical total psychology! But this conclusion must be slightly weakened in consideration of the fact that the input-output specifications of the two subjects realizing the same Turing machine may be different and that the individuation of psychologies may have to be made sensitive to input-output specifications (we return shortly to this point). So let us speak of "isomorphic" psychologies for psychologies that are instances of the same Turing machine *modulo* input-output specification. We then have the following result: On machine functionalism, for two psychological subjects to share even a single mental state, their total psychologies must be isomorphic to each other. Recall Putnam's complaint against the psychoneural identity theory: This theory makes it impossible for both humans and octopuses to be in the same pain state unless they share the same brain state, an unlikely possibility. But it would seem that machine functionalism runs into an exactly identical predicament: For an octopus and a human to be in the same pain state, they must share an isomorphic psychology—an unlikely possibility, to say the least! And for two humans to share a single mental state, they must have an exactly identical total psychology (since the same input-output specification presumably must hold for all or most humans). No analogous consequence follows from the psychoneural identity theory; in this respect, therefore, machine functionalism seems to fare worse than the theory it hopes to replace. All this is a consequence of a fact mentioned earlier, namely, that on functionalism, the individuation of mental kinds is essentially holistic.

Things are perhaps not as bleak for machine functionalism, however, as they might appear, for the following line of response seems available: For both humans and octopuses to be in pain, it is not necessary that *total* octopus psychology coincide with, or be isomorphic to, *total* human psychology. It is only necessary that there be *some* Turing machine that is a correct machine description of both and in which pain figures as an internal machine state; it does not matter if this shared Turing machine falls short of the maximally detailed Turing machines that describe them (these machines represent their "total psychologies"). So what is necessary is that humans and octopuses share a partial, or abbreviated, psychology that encompasses pains. Whether or not pain psychology can so readily be isolated, or abstracted, from the total psychology is a question worth pondering, especially in the context of the functionalist conception of mentality, but there is another potential difficulty here that we should briefly consider.

Recall the point that all this talk of humans' and octopuses' realizing a Turing machine is relative to an input-output specification. Doesn't this mean,

in view of our earlier discussion of a real psychological subject and a computer simulation of one, that the input and output conditions characteristic of humans when they are in pain must be appropriately similar, if not identical, to those characteristic of octopuses' pains, if both humans and octopuses can be said to be in pain? Consider the output side: Do octopuses wince and groan in reaction to pain? They perhaps can wince, but they surely cannot groan or scream and yell "Ouch!" How similar is octopuses' escape behavior, from the purely physical point of view, to the escape behavior of, say, middle-aged, middle-class American males? Is there an abstract enough *nonmental* description of pain behavior that is appropriate for humans and octopuses and all other pain-capable organisms and systems? If there is not, machine functionalism seems to succumb again to the same difficulty that the functionalist has charged against the brain-state theory: An octopus and a human cannot be in the same pain state. Again, the best bet for the functionalist seems to be to appeal to the "teleological appropriateness" of an octopus's and a person's escape behaviors—that is, the fact that the behaviors are biologically appropriate responses to the stimulus conditions in enhancing their chances of survival and their well-being in their respective environments.

There is a further "appropriateness" condition for Turing machines that we must now consider. You will remember our saying that for a machine functionalist, a system has mentality just in case it realizes an "appropriately complex" Turing machine. This proviso is necessary because there are all sorts of simple Turing machines (recall our sample machines) that clearly do not suffice to generate mentality. But how complex is complex enough? What is complexity anyway, and how is it measured? And what kind of complexity is "appropriate" for mentality? These are important but difficult questions, and machine functionalism, unsurprisingly, has not produced detailed general answers to them. What we have, though, is an intriguing proposal, from Turing himself, of a test to determine whether a computing machine can "think." This is the celebrated "Turing test," and we now turn to this proposal.

The Turing Test

Turing's innovative proposal is to bypass these general theoretical questions about appropriateness in favor of a concrete operational test that can evaluate the performance capabilities of computing machines vis-à-vis average humans who, as all sides would agree, are fully mental.[17] The idea is that if machines can do as well as humans on certain appropriate intellectual tasks, then they must be judged no less psychological ("intelligent") than humans.

What, then, are these tasks? Obviously, they must be those that, intuitively, require intelligence and mentality to perform; Turing describes a game, the "imitation game," to test for the presence of these capacities.

The imitation game is played as follows. There are three players: the interrogator, a man, and a woman, with the interrogator segregated from the other two in another room. The man and woman are known only as "X" and "Y" to the interrogator, whose object is to identify which is the man and which is the woman by asking questions via keyboard terminals and monitors. The man's object is to mislead the interrogator into an erroneous identification, whereas the woman's job is to help the interrogator. There are no restrictions on the topics of the questions asked.

Suppose, Turing says, we now replace the man with a computing machine. Now the aim of the interrogator is to find out which is human and which is a machine. The machine is programmed to fool the interrogator into thinking that it is a human. Will the machine do as well as the man in the first game in fooling the interrogator into making wrong guesses? Turing's proposal is that if the machine does as well as the man, then we must credit it with all the intelligence that we would normally confer on a human; it must be judged to possess the full mentality that humans possess.

The gist of Turing's idea can be captured in a simpler test: By asking questions (or just holding a conversation) via keyboard terminals, can we find out whether we are talking to a human or a computing machine? (This is the way the Turing test is now being performed.) If there is a computer that can consistently fool us so that our success in guessing its identity is no better than what could be achieved by random guesses, we must concede, it seems, that this machine has the kind of mentality that we grant to humans. There already are chess-playing computers that would fool most people this way, but only in playing chess: Average chess players would not be able to tell if they are playing a human opponent or a computer. But the Turing test covers all possible areas of human concern: music and poetry, politics and sports, how to fix a leaking faucet or make a soufflé—no holds are barred.

The Turing test is designed to isolate the questions of intelligence and mentality from irrelevant considerations, such as the appearance of the machine (as Turing points out, it does not have to win beauty contests to qualify as a thinker), details of its composition and structure, whether it speaks and moves about like a human, and so on. The test is to focus on a broad range of rational, intellectual capacities and functions. But how good is the test?

Some have pointed out that the test is both too tough and too narrow. Too tough because something does not have to be smart enough to outwit a

human to have mentality or intelligence; in particular, the possession of a language should not be a prerequisite for mentality (think of mute animals). Human intelligence itself encompasses a pretty broad range, and there appears to be no compelling reason to set the minimal threshold of mentality at the level of performance required by the Turing test. The test is perhaps also too narrow in that it seems at best to be a test for the presence of *humanlike* mentality, the kind of intelligence that characterizes humans. Why couldn't there be creatures, or machines, that are intelligent and have a psychology but would fail the Turing test, which, after all, is designed to test whether the computer can fool a *human* interrogator into thinking it is a *human*? Furthermore, it is difficult to see it as a test for the presence of mental states like sensations and perceptions, although it may be an excellent test of broadly intellectual and cognitive capacities (reasoning, memory, and so on). To see something as a full psychological system we must see it in a real-life context, we might argue; we must see it coping with its environment, receiving sensory information from its surroundings, and behaving appropriately in response to it.

Various replies can be attempted to counter these criticisms, but can we say, as Turing himself did, that the Turing test at least provides us with a *sufficient* condition for mentality, although, for the reasons just given, it cannot be considered a necessary condition? If something passes the test, it is at least as smart as we are, and since we have intelligence and mentality, it would be only fair to grant it the same status—or so we might argue. This reasoning seems to presuppose the following thesis:

> *Turing's Thesis.* If two systems are input-output equivalent, they have the same psychological status; in particular, one is mental just in case the other is.

We call it Turing's Thesis because Turing seems to be committed to it. Why is Turing committed to it? Because the Turing test looks only at inputs and outputs: If two computers produce the same output for the same input, for all possible inputs—that is, if they are input-output equivalent—their performance on the Turing test will be exactly identical, and one will be judged to have mentality if and only if the other is. This means that if two Turing machines are correct behavioral descriptions of some system (relative to the same input-output specification), then they satisfy the "appropriateness" condition to the same degree. (Remember that "appropriateness" here refers to appropriateness for mentality.) In this way the general philosophical

stance implicit in Turing's Thesis is more behavioristic than machine-functionalist. For machine functionalism is consistent with the denial of Turing's thesis: It says that input-output equivalence, or behavioral equivalence, is not sufficient to guarantee the same degree of mentality. What arguably follows from machine functionalism is only that systems that realize the same Turing machine—that is, systems for which an identical Turing machine is a correct machine description—enjoy the same degree of mentality.

It appears, then, that Turing's Thesis is mistaken: Internal processing ought to make a difference to mentality. Imagine two machines, each of which does basic arithmetic operations for integers up to 100. Both give correct answers for any input of the form $n + m$, $n \times m$, $n - m$, and $n \div m$ for whole numbers n and m less than or equal to 100. But one of the machines calculates ("figures out") the answer by applying the usual algorithms we use for these operations, whereas the other has a file in which answers are stored for all possible problems of addition, multiplication, subtraction, and division for integers up to 100, and its computation consists in "looking up" the answer for any problem given to it. The second machine is really more like a filing cabinet than a computing machine; it does nothing that we would normally describe as "calculation" or "computation." Neither machine is nearly complex enough to be considered for possible mentality; however, the example should convince us that we need to consider the structure of internal processing, as well as input-output correlations, in deciding whether a given system has mentality.[18] If this is correct, it shows the inadequacy of a purely behavioral test, such as the Turing test, as a criterion of mentality.

So Turing's Thesis seems incorrect: Input-output equivalence does not imply equal mentality. But this does not necessarily invalidate the Turing test, for it may well be that given the inherent richness and complexity of the imitation game, any computing machine that can consistently fool humans—in fact, any machine that is in the ballpark for the competition—has to be running a highly sophisticated, unquestionably "intelligent" program, and there is no real chance that this machine could be operating like a gigantic filing system with a superfast retrieval mechanism.[19]

The "Chinese Room"

John Searle has constructed an intriguing thought-experiment to show that mentality cannot be equated with a computing machine running a program, no matter how complex, "intelligent," and sophisticated it is.[20] Searle invites us to imagine a room—the "Chinese room"—in which someone (say, Searle

himself) who understands no Chinese is confined. He has a set of rules (the "rule book") for systematically transforming strings of symbols to yield further symbol strings. These symbol strings are in fact Chinese expressions, and the transformation rules are purely *formal* in the sense that their application depends solely on the shapes of the symbols involved, not their meanings. So you can apply these rules without knowing any Chinese; all that is required is that you recognize Chinese characters by their shapes. Searle becomes very adept at manipulating Chinese expressions in accordance with the rules given to him (we may suppose that Searle has memorized the whole rule book) so that every time a string of Chinese characters is sent in, Searle goes to work and promptly sends out an appropriate string of Chinese characters. From the perspective of someone outside the room who understands Chinese, the input strings are questions in Chinese and the output strings sent out by Searle are appropriate responses to these questions. The input-output relationships are what we would expect if someone with a genuine understanding of Chinese, instead of Searle, were locked inside the room. And yet Searle does not understand any Chinese, and there is no understanding of Chinese going on anywhere inside the Chinese room. What goes on in the room is only manipulation of symbols on the basis of their shapes, or "syntax," but real understanding involves "semantics," knowing what these symbols represent, or mean. Although Searle's behavior is input-output equivalent to that of a speaker of Chinese, Searle understands no Chinese.

Now, replace Searle with a computer running Searle's rule book as its program. This changes nothing: Both Searle and the computer are syntax-driven machines manipulating strings of symbols according to their syntax. In general, what goes on inside a computer is exactly like what goes on in the Chinese room (with Searle in it): rule-governed manipulations of symbols based on their syntactic shapes. There is no more understanding of Chinese in the computer than there is in the Chinese room. The conclusion to be drawn, Searle argues, is that mentality is more than rule-governed syntactic manipulation of symbols and that there is no way to get semantics—or what the symbols mean or represent—from their syntax. And this means that understanding and other intelligent mental states and activities cannot arise from mere syntactic processes. Computational processes are essentially and exclusively syntactic; they do not depend at all on what the symbols being manipulated might mean or represent, or whether they mean anything at all.

Searle's argument has elicited a large number of critical responses, and just what the argument succeeds in showing remains highly controversial. Although its intuitive appeal and power cannot be denied, we have to be careful in assessing its significance. The appeal of Searle's example may be due, some have argued, to certain misleading assumptions tacitly made in the way he describes what is going on in the Chinese room. We can agree with Searle that input-output equivalence does not constitute psychological equivalence, and that the fact that the Chinese room is input-output equivalent with a speaker of Chinese does not show that the Chinese room, or Searle with his rule book, is a system with a genuine understanding of Chinese. As has already been pointed out, we must attend to internal processing—the kind of program being run—when considering the question of mentality for the system. And here it may be seriously misleading to represent the man locked inside the room as merely "manipulating symbols." Given the sophisticated linguistic processing that has to be performed, we must expect that an extremely sophisticated and highly complex program, something that may far exceed any computer program that has yet been written, will be necessary. It is by no means clear that any human could manage to do what Searle imagines himself to be doing in the Chinese room—that is, short of throwing away the rule book and learning some real Chinese.

Searle has a reply, however: Make the program as complex and intricate as you want, but no amount of syntactic symbol-pushing will generate meaning and understanding. Computation is syntax-driven: As long as the computer is given the same strings of 0s and 1s, it will generate certain further strings of 0s and 1s, no matter what these strings stand for—the prices of bags of potato chips or the addresses of a group of employees or the temperatures of the major cities in New England. The computer would move through exactly the same computational process even if the 0s and 1s meant nothing at all. However, our intentional states, like beliefs and desires, are what they are because they mean, or represent, something. My belief that it is raining outside has the content "it is raining outside," and in virtue of having this content the belief represents, or purports to represent, a specific weather condition in my environment. Suppose this belief causes in me a desire to take an umbrella to work. This causal relation holds in part because the belief and the desire have the particular contents that they have; my belief does not cause a desire to wear my best suit to work, and the belief that it is sunny outside does not cause me to want to take an umbrella with me. Mental processes are driven by representational contents, or meanings. Hence, they cannot be syntax-driven computational processes, and the mind cannot just be a computer running a program, no matter how complex and sophisticated the program may be. The mind is a "semantic engine"; the computer, in contrast, is only a "syntactic engine."

If this is the general argument underlying the Chinese room thought-experiment, it clearly raises a legitimate and perplexing issue about meaning and mental causation. However, this is a general problem that arises for any materialist conception of mentality, quite apart from the specific issue of the computational account of mentality. Consider the position that Searle himself favors: Mentality can arise only in complex biological systems, like the human brain. It seems that the same neurobiological causal processes will go on no matter what the neural states involved represent about the world or whether they represent anything at all. Neural processes seem no more responsive to meaning and representational content than are computational processes. Local physical-biological conditions in the brain, not the distal states of affairs represented by them, are what drive neural processes. If so, isn't Searle in the same boat as Turing and other computationalists?

There is also an important prior question: How do neural states get to represent anything? That is, how do they come to have representational content, and moreover, how do they get to have the particular content that they have? An influential view is that content or meaning arises from our complex interaction with the world around us—in particular, perception and action. So why not make our computer into a robot with a capacity for perception, inference, and action and embed it in the world, like the android Commander Data in one of the *Star Trek* series? (This is what Searle calls the "robot reply"; he rejects it, however.) Wouldn't this also help solve the question of the "teleological aptness" of input-output correlations that we discussed earlier? Our later discussion (in chapter 9) of the general question of how mental states come to have the content they have will be relevant to an assessment of the Chinese room argument.

Further Readings

The classic source of machine functionalism is Hilary Putnam's "Psychological Predicates" (later reprinted as "The Nature of Mental States"). See also his "Robots: Machines or Artificially Created Life?" and "The Mental Life of Some Machines"; all three papers are reprinted in his *Mind, Language, and Reality: Philosophical Papers*, volume 2. The first of these is widely reprinted elsewhere, including *Philosophy of Mind: Classical and Contemporary Readings*, edited by David J. Chalmers, and *Philosophy of Mind: A Guide and Anthology*, edited by John Heil. Ned Block's "What Is Functionalism?" is a clear and concise introduction to functionalism.

For a teleological approach to functionalism, see William G. Lycan, *Consciousness*, chapter 4. For a general biological-evolutionary perspective on mentality, see Ruth G. Millikan, *Language, Thought, and Other Biological Categories*.

For issues involving the Turing test and the Chinese room argument, see Alan M. Turing, "Computing Machinery and Intelligence"; John R. Searle, "Minds, Brains, and Programs"; and Ned Block, "The Mind as Software in the Brain." These articles are reprinted in Heil's *Philosophy of Mind*. Also recommended are Block, "Psychologism and Behaviorism," and Daniel C. Dennett, *Consciousness Explained*, chapter 14. Entries on "Turing Test" and "Chinese Room Argument" in the *Stanford Online Encyclopedia of Philosophy* are useful resources.

For criticisms of machine functionalism (and functionalism in general), see Ned Block, "Troubles with Functionalism," and John R. Searle, *The Rediscovery of the Mind*.

Notes

1. Later retitled "The Nature of Mental States" (1979).

2. Donald Davidson's argument for mental anomalism, as we shall see in chapter 10, also played a part in the decline of reductionism.

3. At least some of them, for it could be argued that certain psychological states can be had only by materially embodied subjects—for example, feelings of hunger and thirst, bodily sensations like pain and itch, and sexual desire.

4. The terms "realize" and "realizer" are given explicit explanations in a later section. In the meantime, you will not go far astray if you read "P realizes M" as "P is a neural substrate, or base, of M."

5. This principle entails mind-body supervenience, which we characterized as minimal physicalism in chapter 1. Further, it arguably entails the thesis of ontological physicalism, as stated in that chapter.

6. See Ronald Melzack, *The Puzzle of Pain*, pp. 15–16.

7. Some have argued that this function-versus-mechanism dichotomy is pervasive at all levels, not restricted to the mental-physical case; see, for example, William G. Lycan, *Consciousness*.

8. As I take it, something like this is the point of Karl Lashley's principle of "equipotentiality"; see his *Brain Mechanisms and Intelligence*, p. 25.

9. To borrow Ned Block's formulation of the question in "What Is Functionalism?" pp. 178–179.

10. As we shall see in connection with machine functionalism, there is another sense of "function," the mathematical sense, involved in "functionalism."

11. Strictly speaking, it is more accurate to say that having *the capacity to sense pain* is being equipped with a tissue-damage detector, and that pain, as an occurrence, is the activation of such a detector.

12. See, for example, B. F. Skinner, "Selections from *Science and Human Behavior*."

13. A treatment of the mathematical theory of computability in terms of Turing machines can be found in Martin Davis, *Computability and Unsolvability*, and in George S. Boolos, John P. Burgess, and Richard C. Jeffrey, *Computability and Logic*.

14. See, for example, Putnam, "Psychological Predicates."

15. For a statement and defense of a position of this kind, see Bas Van Fraassen, *The Scientific Image*.

16. Is there, for any given psychological subject, a *unique* Turing machine that is a machine description (relative to a specification of input and output conditions), or can there be (perhaps there always must be) multiple, nontrivially different machine descriptions? Does realism

about psychology require that there be a unique one? The reader is invited to reflect on these questions.

17. Alan M. Turing, "Computing Machinery and Intelligence."
18. For an elaboration of this point, see Ned Block, "Psychologism and Behaviorism."
19. Daniel C. Dennett, *Consciousness Explained*, pp. 435–440.
20. John R. Searle, "Minds, Brains, and Programs."

6

Mind as a Causal System

Causal-Theoretical Functionalism

In the preceding chapter, we discussed the functionalist attempt to use the concept of a Turing machine 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 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 by receiving sensory inputs and emitting behavioral 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 events, 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 (chapter 5). To achieve the same end, causal-theoretical functionalism attempts to use 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.

The Ramsey-Lewis Method

Consider the following "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 expressions 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. What kind of "theory" T must be if T is to serve as a basis of functional definitions of "pain" and other

mental expressions is a question taken up in a later section. Here T serves only as an example to illustrate the formal technique originally due to Frank Ramsey and adapted by David Lewis for formulating functional definitions of mental kinds.[1]

We first "Ramseify" T by "existentially generalizing" over each mental expression 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 "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 inferential connections between nonpsychological statements.[2] 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:[3]

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

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

x is in distress = $_{def}\exists M_1, M_2, M_3$ [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 like 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 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 much more 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.

The Choice of an Underlying Psychological Theory

So what should the underlying psychological theory T be like if it is to yield, by the Ramsey-Lewis technique, appropriate 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 as one that is distinct from other psychological properties. Given all this, there are two major possibilities to consider.

We might, with Lewis, consider using the platitudes of our shared *commonsense psychology* to form 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.[4] 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.

We must remember that commonsense psychology is, well, only commonsensical: It may be incomplete and partial or contain serious errors. 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*, 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.

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 have nonempty extensions—that is, on there being any instances to which the concepts apply. For if T is false, its Ramseification, T_R, may also be false; in fact, 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, 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 render 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, perhaps 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?

Consider now 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 false and that, considered as a theory, it must be abandoned.[5] 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. Truly massive tissue damage may well 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 domains, it is reasonable to expect that most of what we now consider our best theories in psychology will be abandoned and replaced 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 underwrites the very possibility of making sense of why people do what they do. It is plausible to think that it 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 heroes and heroines 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 clear 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[6]), 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. We could insist that even so, the Ramsey-Lewis method does yield appropriate definitions of these concepts, noting that it is a fact about the concepts of a seriously flawed theory (for example, the concepts of phlogiston, magnetic effluvium, and caloric fluid) that they have no applications in the real world. The difficulty with the Ramsey-Lewis method, however, is that vernacular psychology does not have to be seriously flawed to render its concepts empty; a single false psychological platitude is enough to bankrupt the whole system, by making all of our psychological concepts inapplicable to anything real.

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 psychology, 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_1, you are in pain, and when Mork the Martian is in I_1, he is in pain, since each of you satisfies the functionalist definition of pain as stated.

Functionalism as Physicalism

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 discuss. 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 given 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 psychological mechanism that connects 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 and hence the theory to be Ramseified to yield causal-functional definitions of mental states?

If psychology is a truly autonomous special science, under no methodological, theoretical, or ontological 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 hard, 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 clear, then, that 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 discriminate 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.

A functionalist who accepts (P)—that is, a physicalist functionalist—will interpret the ontology of our original, un-Ramseified psychological theory in an analogous way: The internal states posited by a correct psychological theory are physical-neural states. These considerations have the following implications for psychology: Unless these physicalist constraints are introduced, there is no way of discriminating between behaviorally adequate psychologies; conversely, the 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 must be physically realized.

This appears to reflect the actual research strategies in psychology 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 (for instance, proving logic theorems) 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 divorced from any physical underpinnings; perhaps they are some ghostly mechanisms in Cartesian mental substances. This may be a logically possible position, but hardly a plausible one (see chapter 2). In any case, as noted earlier, physicalism is the default position for discussions in contemporary philosophy of mind.

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 to functionalism generally, and these will not be taken up again here.

Qualia

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. Can there be anything more obvious than that?

Sensations have characteristic *qualitative* features; these are called "phenomenal" or "phenomenological" or "sensory" qualities—"qualia" is now the standard term. Seeing a ripe tomato has a certain distinctive sensory quality that is unmistakably different from the sensory quality involved in seeing a bunch 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 are temporarily unable 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, 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 conceivable: When you look at a ripe tomato, your color experience is like the color I 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 asked what color ripe tomatoes are, and we both describe the color of spinach as "green"; we are equally good at picking tomatoes out of mounds of lettuce leaves. In fact, you and I both seem to be able to imagine that your color spectrum is systematically inverted with respect to mine, without this being manifested in any behavioral difference. 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.[7] 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 physical-neural states, but this supervenience assumption is 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.[8] We return to the issue of qualia in connection with the more general question of consciousness and its reducibility (chapters 8 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 ("nociceptive neurons") somewhere in the brain that gets activated when we experience pain, and similarly for the itch box. When pain sensors 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 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. 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 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 obviously 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 brain-state 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 disembodied brain kept alive in a vat of liquid is maintained in a normal state of consciousness by being fed appropriate 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 mechanism: 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 (that is, 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 many (in fact, indefinitely many) 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 from one species to another.

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, that realizes M on that occasion. Carburetors can have no causal powers beyond those of the physical structures that perform 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).[9] 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?

Let us consider this issue in some detail by reference to machine functionalism. A psychological subject, on this version of functionalism, is a physical-biological system that realizes an appropriate Turing machine (relative to some input-output specification). And for it 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 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.

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, perhaps 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 unique set of causal powers. 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 associated with 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.[10] 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 disqualifies 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 arguably is the received view concerning the nature and status of psychology. But if our reasoning here is at all in the right direction, the conjunction of functionalism and the multiple realizability of the mental apparently 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 required of scientific kinds.[11]

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 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) cognitive and behavioral science is an autonomous science irreducible to more basic, "lower-level" sciences like biology and physics. The fragmentation of psychological-cognitive properties as scientific properties was made plausible, they will argue, 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 these properties represent, and we should never forget that psychological-cognitive properties are "role" properties. So we might want to distinguish between "role functionalism" and "realizer functionalism."[12] Role functionalism identifies each mental property with being in a state that plays a specified causal role (we can call such properties "role" properties) 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.

> *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.

So where the realizer functionalist sees differences and disunity among instances of pain with different realizers, the role functionalist sees similarities and unity. 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.[13]

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.

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.[14] How these empirical results are to be correctly interpreted and understood, however, is an open question. 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. Readers who have a background in these sciences are invited to reflect on the issues further.

Further Readings

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 chapter 8 and the suggested readings therein. On the causal powers of functional properties, see Ned Block, "Can the Mind Change the World?"

The most influential statement of the multiple realization argument is Jerry Fodor, "Special Sciences, or the Disunity of Science as a Working Hypothesis." For discussion and analysis, see Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction." Replying to Kim are Ned Block, "Anti-Reductionism Slaps Back," and Jerry Fodor, "Special Sciences: Still Autonomous After All These Years." For a comprehensive defense of the possibility of cognitive science, see Don Ross and David Spurrett, "What to Say to a Skeptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists."

Notes

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

2. 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."

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

4. 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 behavior of others, using ourselves as models. See Robert M. Gordon, "Folk Psychology as Simulation," and Alvin I. Goldman, "Interpretation Psychologized." 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 determined in greater detail.

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

6. 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 may be possible, however, that there is a core set of principles of vernacular psychology that can be considered a priori true and that suffice as a basis of the application of the Ramsey-Lewis method.

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

8. 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?"

9. 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.

10. See, for example, Jerry Fodor, *Psychosemantics*, ch. 2.

11. For further discussion, see Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction."

12. 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 Ned Block's distinction between the functional-state identity theory and the functional specification theory in his "What Is Functionalism?"

13. Ross and Spurrett, "What to Say to a Skeptical Metaphysician."

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

7

Mental Causation

Causal relations involving mental events are among the familiar facts of everyday experience. My fingers are busily dancing about on the computer keyboard because I want to write about mental causation. The word "because" connecting my want and the movements of my fingers is naturally taken to express a causal connection: My want causes my fingers to move. And we can causally explain the movements of my fingers—for example, why they hit the keys *m*, *e*, *n*, *t*, *a*, and *l* in succession—by reference to my desire to type the word "mental". This is a case of *mental-to-physical* causation. There are two other kinds of causal relations in which mental events figure: *physical-to-mental* and *mental-to-mental* causation. Sensations are among the familiar examples involving causal relations of the physical-to-mental kind: Burns cause pains, irradiations of the retina cause visual sensations, and food poisoning can cause nauseous feelings. Instances of mental-to-mental causation are equally familiar. We often believe one thing (say, that we had better take an umbrella to work) because we believe another thing (say, that it is going to rain later today). This is a case in which a belief causes another belief. Your belief that you have won a fellowship to graduate school causes a feeling of pride and satisfaction, which in turn causes you to want to call your parents. On a grander scale, it is human knowledge, desires, dreams, greed, and ambitions that led our forebears to build the ancient