its skills, emotions, and appetites. Aristotle continued this unlikely dichotomy when he separated the theoretical from the practical, and defined man as a rational animal—as if one could separate man’s rationality from his animal needs and desires. If one thinks of the importance of the sensory-motor skills in the development of our ability to recognize and cope with objects, or of the role of needs and desires in structuring all social situations, or finally of the whole cultural background of human self-interpretation involved in our simply knowing how to pick out and use chairs, the idea that we can simply ignore this know-how while formalizing our intellectual understanding as a complex system of facts and rules is highly implausible.

Great artists have always sensed the truth, stubbornly denied by both philosophers and technologists, that the basis of human intelligence cannot be isolated and explicitly understood. In *Moby-Dick* Melville writes of the tattooed savage, Queequeg, that he had “written out on his body a complete theory of the heavens and the earth, and a mystical treatise on the art of attaining truth; so that Queequeg in his own proper person was a riddle to unfold, a wondrous work in one volume; but whose mysteries not even he himself could read” (1952; p. 477). Yeats puts it even more succinctly: “I have found what I wanted—to put it in a phrase, I say, ‘Man can embody the truth, but he cannot know it.’”

Acknowledgement: I am grateful to John Haugeland for editorial suggestions on transforming this from a book introduction into an independent article.

7

Reductionism and the Nature of Psychology

HILARY PUTNAM

1. Reduction

A DOCTRINE to which most philosophers of science subscribe (and to which I subscribed for many years) is the doctrine that the laws of such “higher-level” sciences as psychology and sociology are reducible to the laws of lower-level sciences—biology, chemistry, ultimately to the laws of elementary particle physics. Acceptance of this doctrine is generally identified with belief in “The Unity of Science” (with capitals), and rejection of it with belief in vitalism, or psychism, or, anyway, something bad.

In this paper I want to argue that this doctrine is wrong. In later sections, I shall specifically discuss the Turing machine model of the mind—and the conception of psychology associated with reductionism and with the Turing machine model. I want to argue that while materialism is right, and while it is true that the only method for gaining knowledge of anything is to rely on testing ideas in practice (and evaluating the results of the tests scientifically), acceptance of these doctrines need not lead to reductionism.

I shall begin with a logical point and then apply it to the special case of psychology. The logical point is that from the fact that the behavior of a system can be deduced from its description as a system of elementary particles, it does not follow that it can be explained from that description. Let us look at an example and then see why this is so.

My example will be a system of two macroscopic objects, a
board in which there are two holes, a square hole one inch across and a round hole one inch in diameter, and a square peg, a fraction less than one inch across. The fact to be explained is: The peg goes through the square hole, and it does not go through the round hole.

One explanation is that the peg is approximately rigid under transportation and the board is approximately rigid. The peg goes through the hole that is large enough and not through the hole that is too small. Notice that the microstructure of the board and the peg is irrelevant to this explanation. All that is necessary is that, whatever the microstructure may be, it be compatible with the fact that the board and the peg are approximately rigid objects.

Suppose, however, we describe the board as a cloud of elementary particles (for simplicity, we will assume these are Newtonian elementary particles) and imagine ourselves given the position and velocity at some arbitrary time \( t_0 \) of each one. We then describe the peg in a similar way. (Say the board is “cloud B” and the peg is “cloud A”.’) Suppose we describe the round hole as “region 1” and the square hole as “region 2”. Let us say that by a heroic feat of calculation we succeed in proving that “cloud A” will pass through “region 2” but not through “region 1”. Have we explained anything?

It seems to me that whatever the pragmatic constraints on explanation may or may not be, one constraint is surely this: The relevant features of a situation should be brought out by an explanation and not buried in a mass of irrelevant information. By this criterion, it seems clear that the first explanation—which points out that the two macro-objects are approximately rigid and that one of the two holes is big enough for the peg and the other is not—explains why “cloud A” passes through “region 2” and never through “region 1”, while the second—the deduction of the fact to be explained from the positions and velocities of the elementary particles, their electrical attractions and repulsions, etc.—fails to explain.

If this seems counterintuitive, it is for two reasons, I think. (1) We have been taught that to deduce a phenomenon in this way is to explain it. But this is ridiculous on the face of it. Suppose I deduce a fact \( F \) from \( G \) and \( I \), where \( G \) is a genuine explanation and \( I \) is something irrelevant. Is \( G \) and \( I \) an explanation of \( F \)?

Normally we would answer, “No. Only the part \( G \) is an explanation”. Now, suppose I subject the statement \( G \) and \( I \) to logical transformations so as to produce a statement \( H \) that is mathematically equivalent to \( G \) and \( I \) (possibly in a complicated way), but such that the information \( G \) is, practically speaking, virtually impossible to recover from \( H \). Then on any reasonable standard the resulting statement \( H \) is not an explanation of \( F \); but \( F \) is deducible from \( H \). I think that the description of the peg and board in terms of the positions and velocities of the elementary particles, their electrical attractions and repulsions, etc., is such a statement \( H \): The relevant information, that the peg and the board are approximately rigid, and the relative sizes of the holes and the peg are buried in this information, but in a useless way (practically speaking). (2) We forget that explanation is not transitive. The microstructure of the board and peg may explain why the board and the peg are rigid, and the rigidity is part of the explanation of the fact that the peg passes through one hole and not the other, but it does not follow that the microstructure, so to speak “raw”—as an assemblage of positions, velocities, etc.—explains the fact that the peg passes through one hole and not the other. Even if the microstructure is not presented “raw”, in this sense, but the information is organized so as to give a revealing account of the rigidity of the macro-objects, a revealing explanation of the rigidity of the macro-objects is not an explanation of something which is explained by that rigidity. If I want to know why the peg passes through one hole and not the other in a normal context (e.g., I already know that these macro-objects are rigid), then the fact that one hole is bigger than the peg all around and the other isn’t is a complete explanation. That the peg and the board consist of atoms arranged in a certain way, and that atoms arranged in that way form a rigid body, etc., might also be an explanation—although one which gives me information (why the board and the peg are rigid) I didn’t ask for. But at least the relevant information—the rigidity of the board and the peg, and the relation of the sizes and shapes of the holes and the pegs—are still explicit. That the peg and the board consist of atoms arranged in a certain way by itself does not explain why the peg goes through one hole and not the other, even if it explains something which in turn explains that.
The relation between (1) and (2) is this: An explanation of an explanation (a “parent” of an explanation, so to speak), generally contains information I, which is irrelevant to what we want to explain, and in addition it contains the information which is relevant, if at all, in a form that may be impossible to recognize. For this reason a parent of an explanation is generally not an explanation.

What follows is that certain systems can have behaviors to which their microstructure is largely irrelevant. For example, a great many facts about rigid bodies can obviously be explained just from their rigidity and the principles of geometry, as in the example just given, without at all going into why those bodies are rigid. A more interesting case is the one in which the higher-level organizational facts on which an explanation turns themselves depend on more than the micro-structure of the body under consideration. This, I shall argue, is the typical case in the domain of social phenomena.

For an example, consider the explanation of social phenomena. Marx in his analysis of capitalism uses certain facts about human beings—for example, that they have to eat in order to live, and they have to produce in order to eat. He discusses how, under certain conditions, human production can lead to the institution of exchange, and how that exchange in turn leads to a new form of production, production of commodities. He discusses how production of commodities for exchange can lead to production of commodities for profit, to wage labor and capital.

Assume that something like this is right. How much is the microstructure of human beings relevant? The case is similar to the first example in that the specifics or the microstructure are irrelevant: What is relevant is, so to speak, an organizational result of microstructure. In the first case the relevant organizational result was rigidity: In the present case, the relevant organizational result is intelligent beings able to modify both the forces of production and the relations of production to satisfy both their basic biological needs and those needs which result from the relations of production they develop. To explain how the microstructure of the human brain and nervous system accounts for this intelligence would be a great feat for biology; it might or might not have relevance for political economy.

But there is an important difference between the two examples.

Given the micro-structure of the peg and the board, one can deduce the rigidity. But given the micro-structure of the brain and the nervous system, one cannot deduce that capitalist production relations will exist. The same creatures can exist in precapitalist commodity production, or in feudalism, or in socialism, or in other ways. The laws of capitalist society cannot be deduced from the laws of physics plus the description of the human brain: They depend on “boundary conditions” which are accidental from the point of view of physics but essential to the description of a situation as “capitalism”. In short, the laws of capitalism have a certain autonomy vis-à-vis the laws of physics: They have a physical basis (men have to eat), but they cannot be deduced from the laws of physics. They are compatible with the laws of physics; but so are the laws of socialism and of feudalism.

This same autonomy of the higher-level science appears already at the level of biology. The laws that collectively make up the theory of evolution are not deducible from the laws of physics and chemistry; from the latter laws it does not even follow that one living thing will live for five seconds, let alone that living things will live long enough to evolve. Evolution depends on a result of microstructure (variation in genotype); but it also depends on conditions (presence of oxygen, etc.) which are accidental from the point of view of physics and chemistry. The laws of the higher-level discipline are deducible from the laws of the lower-level discipline together with “auxiliary hypotheses” which are accidental from the point of view of the lower-level discipline. And most of the structure at the level of physics is irrelevant from the point of view of the higher-level discipline; only certain features of that structure (variation in genotype, or rigidity, or production for profit are relevant), and these are specified by the higher-level discipline, not the lower-level one.

The alternative, mechanism or vitalism, is a false alternative. The laws of human sociology and psychology, for example, have a basis in the material organization of persons and things, but they also have the autonomy just described vis-à-vis the laws of physics and chemistry.

The reductionist way of looking at science both springs from and reinforces a specific set of ideas about the social sciences. Thus human biology is relatively unchanging. If the laws of psychology
are deducible from the laws of biology and (also unchanging) reductive definitions, then it follows that the laws of psychology are also unchanging. Thus the idea of an unchanging human nature—a set of structured psychological laws, dependent on biology but independent of sociology—is presupposed at the outset. Also, each science in the familiar sequence—physics, chemistry, biology, psychology, sociology—is supposed to reduce to the one below (and ultimately to physics). Thus sociology is supposed to reduce to psychology which in turn reduces to biology via the theory of the brain and nervous system. This assumes a definite attitude toward sociology, the attitude of methodological individualism. (In conventional economics, for example, the standard attitude is that the market is shaped by the desires and preferences of individual people, no conceptual apparatus even exists for investigating the ways in which the desires and preferences of individuals are shaped by the economic institutions.)

Besides supporting the idea of an unchanging human nature and methodological individualism, there is another and more subtle role that reductionism plays in one’s outlook. This role may be illustrated by the effect of reductionism on biology departments: When Crick and Watson made their famous discoveries, many biology departments fired some or all of their naturalists! Of course, this was a crude mistake. Even from an extreme reductionist point of view, the possibility of explaining the behavior of species via DNA mechanisms “in principle” is very different from being able to do it in practice. Firing someone who has a lot of knowledge about the habits of, say, bats, because someone else has a lot of knowledge about DNA is a big mistake. Moreover, as we saw above, you can’t explain the behavior of bats, or whatever species, just in terms of DNA mechanisms—you have to know the “boundary conditions”. That a given structure enables an organism to fly, for example, is a function not just of its strength, etc., but also of the density of the earth’s atmosphere. And DNA mechanisms represent the wrong level of organization of the data—what one wants to know about the bat, for example, is that it has mechanisms for producing supersonic sounds, and mechanisms for “triangulating” on its own reflected sounds (“echolocating”).

The point is that reductionism comes on as a doctrine that breeds respect for science and the scientific method. In fact, what it breeds is physics worship coupled with neglect of the “higher-level” sciences. Infatuation with what is supposedly possible “in principle” goes with indifference to practice and to the actual structure of practice.

I don’t mean to ascribe to reductionists the doctrine that the “higher-level” laws could be arrived at in the first place by deduction from the “lower-level” laws. Reductionist philosophers would very likely have said that firing the naturalist was a misapplication of their doctrines, and that neglect of direct investigation at “the level of sociology” would also be a misapplication of their doctrine. What I think goes on is this: Their claim that higher-level laws are deducible from lower-level laws and therefore higher-level laws are explainable by lower-level laws involves a mistake (in fact, two mistakes). It involves neglect of the structure of the higher-level explanations which reductionists never talk about at all, and it involves neglect of the fact that more than one higher-level structure can be realized by the lower-level entities (so that what the higher-level laws are cannot be deduced from just the laws obeyed by the “lower-level” entities). Neglect of the “higher-level” sciences themselves seems to me to be the inevitable corollary of neglecting the structure of the explanations in those sciences.

2. Turing Machines

In previous papers (1960; 1965; 1967), I have argued for the hypothesis that (1) a whole human being is a Turing machine, and (2) that psychological states of a human being are Turing machine states or disjunctions of Turing machine states. In this section I want to argue that this point of view was essentially wrong, and that I was too much in the grip of the reductionist outlook just described.

Let me begin with a technical difficulty. A state of a Turing machine is described in such a way that the machine can be in exactly one state at a time. Moreover, memory and learning are not represented in the Turing machine model as acquisition of new states, but as acquisition of new information printed on the machine’s tape. Thus if human beings have any states at all which resemble Turing machine states, those states must (1) be states the human can be in at any time, independently of learning and
memory; and (2) be total instantaneous states of the human being—states which determine, together with learning and memory, what the next state will be, as well as totally specifying the present condition of the human being ("totally" from the standpoint of psychological theory, that means).

These characteristics already establish that no psychological state in any customary sense can be a Turing machine state. Take a particular kind of pain to be a "psychological state". If I am a Turing machine, then my present "state" must determine not only whether I am having that particular kind of pain, but also whether I am about to say "three", whether I am hearing a shrill whine, etc. So the psychological state in question (the pain) is not the same as my "state" in the sense of machine state, although it is possible (so far) that my machine state determines my psychological state. Moreover, no psychological theory would pretend that having a pain of a particular kind, being about to say "three", or hearing a shrill whine, etc. all belong to one psychological state, although there could well be a machine state characterized by the fact that I was in it only when simultaneously having that pain, being about to say "three", hearing a shrill whine, etc. So, even if I am a Turing machine, my machine states are not the same as my psychological states. My description qua Turing machine (machine table) and my description qua human being (via a psychological theory) are descriptions at two totally different levels of organization.

So far it is still possible that a psychological state is a large disjunction (practically speaking, an almost infinite disjunction) of machine states, although no single machine state is a psychological state. But this is very unlikely when we move away from states like "pain" (which are almost biological to states like "jealousy" or "love" or "competitiveness". Being jealous is certainly not an instantaneous state, and it depends on a great deal of information and on many learned facts and habits. But Turing machine states are instantaneous and are independent of learning and memory. That is, learning and memory may cause a Turing machine to go into a state, but the identity of the state does not depend on

---

1 For an exposition of Turing machines, see Davis (1958); there is also an attractive little monograph by Trachtenbrot on the subject.
Hilary Putnam

Reductionism and the Nature of Psychology

compatible with the Turing machine model, with the physicalist model, etc. It is in this attenuated sense that the laws of any higher-level discipline have to be compatible with the laws of physics: it has to be compatible with the laws of physics that the higher-level laws could be true to within the required accuracy. But the model associated with the higher-level laws need not at all be compatible with the model associated with the lower-level laws. Another way of putting the same point is this: Let L be the higher-level laws as normally stated in psychology texts (or texts of political economy, or whatever). Let L* be the statement “L is approximately correct.” Then it is only L* that has to be compatible with the laws of physics, not L. (2) The exact law has to be compatible with the Turing machine model (or anyway the laws of physics). False. There need not be any “exact” law—any law more exact than Hull’s—at the psychological level. In each individual case of rote learning, the exact description of what happened has to be compatible with the laws of physics. But the best statement one can make in the general case, at the psychological level of organization, may well be that Hull’s laws are correct to within random errors whose explanation is beneath the level of psychology.

The general picture, it seems to me, is this: Each science describes a set of structures in a somewhat idealized way. It is sometimes believed that a non-idealized description, an “exact” description, is possible “in principle” at the level of physics; be that as it may, there is not the slightest reason to believe that it is possible at the level of psychology or sociology. The difference is this: If a model of a physical structure is not perfect, we can argue that it is the business of physics to account for the inaccuracies. But if a model of a social structure is not perfect, if there are unsystematic errors in its application, the business of accounting for those errors may or may not be the business of social science. If a model of, say, memory in functional terms (e.g., a flow chart for an algorithm) fails to account for certain memory losses, that may be a better psychological theory of memory (different flow chart) is called for, or because on certain occasions memory losses are to be accounted for by biology (an accident in the brain, say) rather than by psychology.

If this picture is correct, then “oversimplified” models may well be the best possible at the “higher” levels. And the relationship to physics is just this: It is compatible with physics that the “good” models on the higher levels should be approximately realized by systems having the physical constitution that human beings actually have.

At this point, I should like to discuss an argument proposed by Hubert Dreyfus. Dreyfus believes that the functional organization of the brain is not correctly represented by the model of a digital computer. As an alternative he has suggested that the brain may function more like an analog computer (or a complicated system of analog computers). One kind of analog computer mentioned by Dreyfus is the following: Construct a map of the railway system of the U.S. made out of string (the strings represent the railroad lines, the knots represent the junctions). Then to find the shortest path between any two junctions—say, Miami and Las Vegas—just pick up the map by the two corresponding knots and pull the two knots away from each other until a string between them becomes straight. That string will represent the shortest path.

When Dreyfus advanced this in conversation, I rejected it on the following grounds: I said that the physical analog computer (the map) really was a digital computer, or could be treated as one, on the grounds that (1) matter is atomic; (2) one could treat the molecules of which the string consists as gears which are capable of assuming a discrete number of positions vis-à-vis adjacent molecules. Of course, this only says that the analog computer can be well approximated by a system which is digital. What I overlooked is that the atomic structure of the string is irrelevant to the working of the analog computer. Worse, I had to invent a microstructure which is just as fictitious as the idealization of a continuous string of constant length (the idea of treating molecules as gears) in order to carry through the re-description of the analog device as a digital device. The difference between my idealization (strings of gears) and the classical idealization (continuous strings) is that the classical idealization is relevant to the functioning of the device as an analog computer (the device works because the strings are—approximately—continuous strings), while my idealization is irrelevant to the description of the system on any level.
3. Psychology

The previous considerations show that the Turing machine model need not be taken seriously as a model of the functional organization of the brain. Of course, the brain has digital elements—the yes-no firing of the neurons—but whether the larger organization is correctly represented by something like a flow chart for an algorithm or by something quite different we have no way of knowing right now. And Hull's model for rote learning suggests that some brain processes are best conceptualized in terms of continuous rather than discrete variables.

In the first section of this paper we argued that psychology need not be deducible from the laws describing the functional organization of the human brain, and in the previous section we used a psychological state (jealousy) to illustrate that the Turing machine model cannot be correct as a paradigm for psychological theory.

In short, there are two different questions which have got confused in the literature: (1) Is the Turing machine model correct as a model for the functional organization of the human brain? and (2) Is the Turing machine model correct as a model for psychological theory? Only on the reductionist assumption that psychology is the description of the functional organization of the brain, or something very close to it, can these two questions be identified.

Our answer to these two questions so far is that (1) there is little evidence that the Turing machine model is correct as a model of the functional organization of the brain; and (2) the Turing machine model cannot be correct as a model for psychological theory—i.e., psychological states are not machine states nor are they disjunctions of machine states. But what is the nature of psychological states?

The idea of a fixed repertoire of emotions, attitudes, etc., independent of culture is easily seen to be questionable. An attitude that we are very familiar with, for example, is the particular kind of arrogance that one person feels toward other people “because” he does mental work and they do manual work. (The reason I put “because” in shudder quotes is that really the causality is much more complicated—he feels arrogant because his society has successfully won him and millions of other people to the idea that the worker is superior to the extent his work differs from the work of a common laborer and resembles that of a manager, perhaps, or because it has won him and millions of other people to the idea that certain kinds of work are inherently above most people—“they couldn’t understand”—etc.). An attitude we find it almost impossible to imagine is the following: One person feeling superior to others because the first person cleans latrines and the others do not. This is not the case because people who clean latrines are innately inferior, not because latrine-cleaning is inherently degrading. Given the right social setting, this attitude which we cannot now imagine would be commonplace. Not only the particular attitudes and emotions we feel are culture-bound, but so are the connections. For example, in our society, arrogance of mental workers is associated with extreme competitiveness; but in a different society it might be associated with the attitude that one is above competing, while being no less arrogant. This might be a reflection of the difference between living in a society based on competition and living in a society based on a feudal hierarchy.

Anthropological literature is replete with examples that support the idea that emotions and attitudes are culture-dependent. For example, there have existed and still exist cultures in which private property and the division of labor are unknown. An Arunha cannot imagine the precise attitude with which Marie Antoinette said “Let them eat cake”, nor the precise attitude of Richard Herrnstein toward the “residue” of low I.Q. people which, he says, is being “precipitated”, nor the precise attitude which made me and thousands of other philosophers feel tremendous admiration toward John Austin for distinguishing “Three Ways of Spilling Ink” (“intentionally”, “deliberately”, and “on purpose”). Nor can we imagine many of the attitudes which Arunha feel, and which are bound up with their culture and religion.

This suggests the following thesis: Psychology is as under-determined by biology as it is by elementary particle physics, and people's psychology is partly a reflection of deeply entrenched societal beliefs. One advantage of this position is that it permits one to deny that there is a fixed human nature at the level of psychology, without denying that *homo sapiens* is a natural kind
at the level of biology. Marx's thesis that there is no fixed 'human nature' which people have under all forms of social organization was not a thesis about "nature versus nurture."

[Because of its length, Section 4 has been omitted from this edition. It is a discussion of the social and political biases built into the concept of intelligence, including the technical concept of IQ.]

5. Psychology Again

If these reflections are right, then it is worthwhile re-examining the nature of psychology. Reductionism asserts that psychology is deducible from the functional organization of the brain. The foregoing remarks suggest that psychology is strongly determined by sociology. Which is right?

The answer, I suspect, is that it depends on what you mean by psychology. Chomsky remarks that "so far as we know, animals learn according to a genetically determined program." While scientific knowledge reflects the development of a socially determined program for learning, there can be little doubt that the possible forms of socially determined programs must in some ways be conditioned by the "genetically determined program" and presuppose the existence of this program in the individual members of the society. The determination of the truth of this hypothesis and the spelling out of the details are the tasks of cognitive psychology. Nothing said here is meant to downgrade the importance of that task, or to downgrade the importance of determining the functional organization of the brain. Some parts of psychology are extremely close to biology—Hull's work on rote learning, much of the work on reinforcement, and so on. It is no accident that in my own reductionist papers the example of a psychological state was usually "pain," a state that is strongly biologically marked. On the other hand, if one thinks of the parts of psychology that philosophers and clinical psychologists tend to talk about—psychological theories of aggression, for instance, or theories of intelligence, or theories of sexuality, then it seems to me that one is thinking of the parts of psychology which study mainly societal beliefs and their effects in individual behavior.

That these two sides of psychology are not distinguished very clearly is itself an effect of reductionism. If they were, one might have noticed that none of the literature on intelligence in the past seventy-five years has anything in the slightest to do with illuminating the nature and structure of human cognitive capacity.