6
Philosophy and our Mental Life

Hilary Putnam

The question which troubles laymen, and which has long troubled philosophers, even if it is somewhat disguised by today's analytic style of writing philosophy, is this: are we made of matter or soul-stuff? To put it as bluntly as possible, are we just material beings, or are we "something more"? In this paper, I will argue as strongly as possible that this whole question rests on false assumptions. My purpose is not to dismiss the question, however, so much as to speak to the real concern which is behind the question. The real concern is, I believe, with the autonomy of our mental life.

People are worried that we may be debunked, that our behavior may be exposed as really explained by something mechanical. Not, to be sure, mechanical in the old sense of cogs and pulleys, but in the newer sense of electricity and magnetism and quantum chemistry and so forth. In this paper, part of what I want to do is to argue that this can't happen. Mentality is a real and autonomous feature of our world.

But even more important, at least in my feeling, is the fact that this whole question has nothing to do with our substance. Strange as it may seem to common sense and to sophisticated intuition alike, the question of the autonomy of our mental life does not hinge on and has nothing to do with that all too popular, all too old question about matter or soul-stuff. We could be made of Swiss cheese and it wouldn't matter.

Failure to see this, stubborn insistence on formulating the question as *matter or soul,* utterly prevents progress on these questions. Conversely, once we see that our substance is not the issue, I do not see how we can help but make progress.

The concept which is key to unravelling the mysteries in the philosophy of mind, I think, is the concept of *functional isomorphism*. Two systems are functionally isomorphic if *there is a correspondence between the states of one and the states of the other that preserves functional relations*. To start with computing machine examples, if the functional relations are just sequence relations, e.g. *state A is always followed by state B,* then, for *F* to be a functional isomorphism, it must be the case that state *A* is followed by state *B* in system 1 if and only if state *F(A)* is followed by state *F(B)* in system 2. If the functional

relations are, say, data or print-out relations, e.g. *when print π is printed on the tape, system 1 goes into state A,* these must be preserved. *When print π is printed on the tape, system 2 goes into state F(A),* if *F* is a functional isomorphism between system 1 and system 2. More generally, if *T* is a correct theory of the functioning of system 1, at the functional or psychological level, then an isomorphism between system 1 and system 2 must map each property and relation defined in system 2 in such a way that T comes out true when all references to system 1 are replaced by references to system 2, and all property and relation symbols in *T* are reinterpreted according to the mapping.

The difficulty with the notion of functional isomorphism is that it *presupposes the notion of a thing's being a functional or psychological description*. It is for this reason that, in various papers on this subject, I introduced and explained the notion in terms of Turing machines. And I felt constrained, therefore, to defend the thesis that *we* are Turing machines. Turing machines come, so to speak, with a normal form for their functional description, the so-called machine table—a standard style of program. But it does not seem fatally sloppy to me, although it is sloppy, if we apply the notion of functional isomorphism to systems for which we have no detailed idea at present what the normal form description would look like—systems like ourselves. The point is that even if we don't have any idea what a comprehensive psychological theory would look like, I claim that we know enough (and here analogies from computing machines, economic systems, games and so forth are helpful) to point out illuminating differences between any possible psychological theory of a human being, or even a functional description of a computing machine or an economic system, and a physical or chemical description. Indeed, Dennett and Fodor have done a great deal along these lines in recent books.

This brings me back to the question of *copper, cheese, or soul*. One point we can make immediately as soon as we have the basic concept of functional isomorphism is this: two systems can have quite different constitutions and be functionally isomorphic. For example, a computer made of electrical components can be isomorphic to one made of cogs and wheels. In other words, for each state in the first computer there is a corresponding state in the other, and, as we said before, the sequential relations are the same—if state *S* is followed by state *B* in the case of the electronic computer, state *A* would be followed by state *B* in the case of the computer made of cogs and wheels, and it doesn't matter at all that the *physical realizations* of those states are totally different. So a computer made of electrical components can be isomorphic to one made of cogs and wheels or to human clerks using paper and pencil. A computer made of one sort of wire, say copper wire, or one sort of relay, etc. will be in a different physical and chemical state when it computes than a computer made of a different sort of wire and relay. But the functional description may be the same.

We can extend this point still further. Assume that one thesis of materialism (I shall call it the "first thesis") is correct, and we are, as wholes, just material systems obeying physical laws. Then the second thesis of classical materialism

cannot be correct—namely, our mental states, e.g. *thinking about next summer's vacation,* cannot be *identical* with any physical or chemical states. For it is clear from what we already know about computers etc., that whatever the program of the brain may be, it must be physically possible, though not necessarily feasible, to produce something with that same program but quite a different physical and chemical constitution. Then to identify the state in question with its physical or chemical realization would be quite absurd, given that that realization is in a sense quite accidental, from the point of view of psychology, anyway (which is the relevant science).[1] It is as if we met Martians and discovered that they were in all functional respects isomorphic to us, but we refused to admit that they could feel pain because their *C* fibers were different.

Now, imagine two possible universes, perhaps "parallel worlds", in the science fiction sense, in one of which people have good old fashioned souls, operating through pineal glands, perhaps, and in the other of which they have complicated brains. And suppose that the souls in the soul world are functionally isomorphic to the brains in the brain world. Is there any more sense to attaching importance to this difference than to the difference between copper wires and some other wires in the computer? Does it matter that the soul people have, so to speak, immaterial brains, and that the brain people have material souls? What matters is the common structure, the theory *T* of which we are, alas, in deep ignorance, and not the hardware, be it ever so ethereal.

One may raise various objections to what I have said. I shall try to reply to some of them.

One might, for example, say that if the souls of the soul people are isomorphic to the brains of the brain people, then their souls must be automata-like, and that's not the sort of soul we are interested in. "All your argument really shows is that there is no need to distinguish between a brain and an automaton-like soul." But what precisely does that objection come to?

I think there are two ways of understanding it. It might come to the claim that the notion of functional organization or functional isomorphism only makes sense for automata. But that is totally false. Sloppy as our notions are at present, we at least know this much, as Jerry Fodor has emphasized: we know that the notion of functional organization applies to anything to which the notion of a psychological theory applies. I explained the most general notion of functional isomorphism by saying that two systems are functionally isomorphic if there is an isomorphism that makes both of them models for the same psychological theory. (That is stronger than just saying that they are both models for the same psychological theory—they are isomorphic realizations of the same abstract structure.) To say that real old fashioned souls would not be in the domain of definition of the concept of functional organization or of the concept of functional isomorphisms would be to take the position that whatever we mean by the soul, it is something for which there can be no theory. That seems pure obscurantism. I will assume, henceforth, that it is not built into the notion of

mind or soul or whatever that it is unintelligible or that there couldn't be a theory of it.

Secondly, someone might say more seriously that even if there is a theory of the soul or mind, the soul, at least in the full, rich old fashioned sense, is supposed to have powers that no mechanical system could have. In the latter part of this chapter I shall consider this claim.

If it is built into one's notions of the soul that the soul can do things that violate the laws of physics, then I admit I am stumped. There cannot be a soul which is isomorphic to a brain, if the soul can read the future clairvoyantly, in a way that is not in any way explainable by physical law. On the other hand, if one is interested in more modest forms of magic like telepathy, it seems to me that there is no reason in principle why we couldn't construct a device which would project subvocalized thoughts from one brain to another. As to reincarnation, if we are, as I am urging, a certain kind of functional structure (my identity is, as it were, my functional structure), there seems to be in principle no reason why that could not be reproduced after a thousand years or a million years or a billion years. Resurrection: as you know, Christians believe in resurrection in the flesh, which completely bypasses the need for an immaterial vehicle. So even if one is interested in those questions (and they are not my concern in this paper, although I am concerned to speak to people who have those concerns), even then one doesn't need an immaterial brain or soul-stuff.

So if I am right, and the question of matter or soul-stuff is really irrelevant to any question of philosophical or religious significance, why so much attention to it, why so much heat? The crux of the matter seems to be that both the Diderots of this world and the Descartes's of this world have agreed that if we are matter, then there is a physical explanation for how we behave, disappointing or exciting. I think the traditional dualist says *"wouldn't it be terrible if we turned out to be just matter, for then there is a physical explanation for everything we do"*. And the traditional materialist says *"if we are just matter, then there is a physical explanation for everything we do. Isn't that exciting!"* (It is like the distinction between the optimist and the pessimist: an optimist is a person who says "this is the best of all possible worlds"; and a pessimist is a person who says "you're right".)[2]

I think they are both wrong. I think Diderot and Descartes were both wrong in assuming that if we are matter, or our souls are material, then there is a physical explanation for our behavior.

Let me try to illustrate what I mean by a very simple analogy. Suppose we have a very simple physical system—a board in which there are two holes, a circle one inch in diameter and a square one inch high, and a cubical peg one-sixteenth of an inch less than one inch high. We have the following very simple fact to explain: *the peg passes through the square hole, and it does not pass through the round hole*.

In explanation of this, one might attempt the following. One might say that the peg is, after all, a cloud or, better, a rigid lattice of atoms. One might even

attempt to give a description of that lattice, compute its electrical potential energy, worry about why it does not collapse, produce some quantum mechanics to explain why it is stable, etc. The board is also a lattice of atoms. I will call the peg "system *A*", and the holes "region 1" and "region 2". One could compute all possible trajectories of system *A* (there are, by the way, very serious questions about these computations, their effectiveness, feasibility, and so on, but let us assume this), and perhaps one could deduce from just the laws of particle mechanics or quantum electrodynamics that system *A* never passes through region 1, but that there is at least one trajectory which enables it to pass through region 2. Is this an explanation of the fact that the peg passes through the square hole and not the round hole?

Very often we are told that if something is made of matter, its behavior must have a physical explanation. And the argument is that if it is made of matter (and we make a lot of assumptions), then there should be a deduction of its behavior from its material structure. *What makes you call this deduction an explanation?*

On the other hand, if you are not "hipped" on the idea that *the* explanation must be at the level of the ultimate constituents, and that in fact the explanation might have the property that *the ultimate constituents don't matter,* that *only the higher level structure matters,* then there is a very simple explanation here. The explanation is that the board is rigid, the peg is rigid, and as a matter of geometrical fact, the round hole is smaller than the peg, the square hold is bigger than the cross-section of the peg. The peg passes through the hole that is large enough to take its cross-section, and does not pass through the hole that is too small to take its cross-section. That is a correct explanation whether the peg consists of molecules, or continuous rigid substance, or whatever. (If one wanted to amplify the explanation, one might point out the geometrical fact that a square one inch high is bigger than a circle one inch across.)

Now, one can say that in this explanation certain *relevant structural features of the situation* are brought out. The geometrical features are brought out. It is *relevant* that a square one inch high is bigger than a circle one inch around. And the relationship between the size and shape of the peg and the size and shape of the holes is *relevant*. It is *relevant* that both the board and the peg are *rigid* under transportation. And nothing else is relevant. The same explanation will go in any world (whatever the microstructure) in which those *higher level structural features* are present. In that sense *this explanation is autonomous*.

People have argued that I am wrong to say that the microstructural deduction is not an explanation. I think that in terms of the *purposes for which we use the notion of explanation,* it is not an explanation. If you want to, let us say that the deduction *is* an explanation, it is just a terrible explanation, and why look for terrible explanations when good ones are available?

Goodness is not a subjective matter. Even if one agrees with the positivists who saddled us with the notion of explanation as deduction from laws, one of the things we do in science is to look for laws. Explanation is superior not just subjectively, but *methodologically,* in terms of facilitating the aims of scientific

inquiry, if it brings out relevant laws. An explanation is superior if it is more general.

Just taking those two features, and there are many many more one could think of, compare the explanation at the higher level of this phenomenon with the atomic explanation. The explanation at the higher level brings out the relevant geometrical relationships. The lower level explanation conceals those laws. Also notice that the higher level explanation applies to a much more interesting class of systems (of course that has to do with what we are interested in).

The fact is that we are much more interested in generalizing to other structures which are rigid and have various geometrical relations, than we are in generalizing to *the next peg that has exactly this molecular structure,* for the very good reason that there is not going to *be* a next peg that has exactly this molecular structure. So in terms of real life disciplines, real life ways of slicing up scientific problems, the higher level explanation is far more general, which is why it is *explanatory*.

We were only able to deduce a statement which is lawful at the *higher* level, that the peg goes through the hole which is larger than the cross-section of the peg. When we try to deduce the possible trajectories of "system *A*" from statements about the individual atoms, we use premises which are totally accidental —this atom is here, this carbon atom is there, and so forth. And that is one reason that it is very misleading to talk about a reduction of a science like economics to the level of the elementary particles making up the players of the economic game. In fact, their motions—buying this, selling that, arriving at an equilibrium price—these motions cannot be deduced from just the equations of motion. Otherwise they would be *physically necessitated,* not *economically necessitated,* to arrive at an equilibrium price. They play that game because they are particular systems with particular boundary conditions which are totally accidental from the point of view of physics. This means that the derivation of the laws of economics from *just* the laws of physics is *in principle* impossible. The derivation of the laws of economics from the laws of physics and *accidental statements about which particles were where when* by a Laplacian supermind might be in principle possible, but why want it? A few chapters of, e.g. von Neumann, will tell one far more about regularities at the level of economic structure than such a deduction ever could.

The conclusion I want to draw from this is that we do have the kind of autonomy that we are looking for in the mental realm. Whatever our mental functioning may be, there seems to be no serious reason to believe that it is *explainable* by our physics and chemistry. And what we are interested in is not: given that we consist of such and such particles, could someone have predicted that we would have this mental functioning? because such a prediction is not *explanatory,* however great a feat it may be. What we are interested in is: can we say at this autonomous level that since we have this sort of structure, this sort of program, it follows that we will be able to learn this, we will tend to like that,

and so on? These are the problems of mental life—the description of this autonomous level of mental functioning—and that is what is to be discovered.

In previous papers, 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.

Let me begin with a technical difficulty. A *state* of a Turing machine is described in such a way that a Turing 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 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 or not I am having that particular kind of pain, but also whether or not I am about to say "three", whether or not 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, 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

learning and memory, whereas, no matter what state I am in, identifying that state as "being jealous of *X*'s regard for *Y*" involves specifying that I have learned that *X* and *Y* are persons and a good deal about social relations among persons. Thus jealousy can neither be a machine state nor a disjunction of machine states.

One might attempt to modify the theory by saying that being jealous= either being in State *A* and having tape c_1 *or* being in State *A* and having tape c_2 *or*…being in State *B* and having tape d_1 or being in State *B* and having tape d_2…being in State *Z* and having tape y_1…or being in State *Z* and having tape y_n—i.e. define a psychological state as disjunction, the individual disjuncts being not Turing machine states, as before, but conjunctions of a machine state and a tape (i.e. a total description of the content of the memory bank). Besides the fact that such a description would be literally infinite, the theory is now without content, for the original purpose was to use the machine table as a model of a psychological theory, whereas it is now clear that the machine table description, although different from the description at the elementary particle level, is as removed from the description *via* a psychological theory as the physicochemical description is.

What is the importance of machines in the philosophy of mind? I think that machines have both a positive and a negative importance. The positive importance of machines was that it was in connection with machines, computing machines in particular, that the notion of functional organization first appeared. Machines forced us to distinguish between an abstract structure and its concrete realization. Not that that distinction came into the world for the first time with machines. But in the case of computing machines, we could not avoid rubbing our noses against the fact that what we had to count as to all intents and purposes the same structure could be realized in a bewildering variety of different ways; that the important properties were not physical-chemical. That the machines made us catch on to the idea of functional organization is extremely important. The negative importance of machines, however, is that they tempt us to oversimplification. The notion of functional organization became clear to us through systems with a very restricted, very specific functional organization. So the temptation is present to assume that we must have that restricted and specific kind of functional organization.

Now I want to consider an example—an example which may seem remote from what we have been talking about, but which may help. This is not an example from the philosophy of mind at all. Consider the following fact. The earth does not go around the sun in a circle, as was once believed, it goes around the sun in an ellipse, with the sun at one of the foci, not in the center of the ellipse. Yet one statement which would hold true if the orbit was a circle and the sun was at the centre still holds true, surprisingly. That is the following statement: the radius vector from the sun to the earth sweeps out equal areas in equal times. If the orbit were a circle, and the earth were moving with a constant velocity, that would be trivial. But the orbit is not a circle. Also the velocity is not constant—when the earth is farthest away from the sun, it is going most slowly,

when it is closest to the sun, it is going fastest. The earth is speeding up and slowing down. But the earth's radius vector sweeps out equal areas in equal times.[3] Newton deduced that law in his *Principia,* and his deduction shows that the only thing on which that law depends is that the force acting on the earth is in the direction of the sun. That is absolutely the only fact one needs to deduce that law. Mathematically it is equivalent to that law.[4] That is all well and good when the gravitational law is that every body attracts every other body according to an inverse square law, because then there is always a force on the earth in the direction of the sun. If we assume that we can neglect all the other bodies, that their influence is slight, then that is all we need, and we can use Newton's proof, or a more modern, simpler proof.

But today we have very complicated laws of gravitation. First of all, we say what is really going on is that the world lines of freely falling bodies in space-time are geodesics. And the geometry is determined by the mass-energy tensor, and the ankle bone is connected to the leg bone, etc. So, one might ask, how would a modern relativity theorist explain Kepler's law? He would explain it very simply. *Kepler's laws are true because Newton's laws are approximately true.* And, in fact, an attempt to replace that argument by a deduction of Kepler's laws from the field equations would be regarded as almost as ridiculous (but not quite) as trying to deduce that the peg will go through one hole and not the other from the positions and velocities of the individual atoms.

I want to draw the philosophical conclusion that Newton's laws *have a kind of reality in our world* even though they are not *true*. The point is that it will be necessary to appeal to Newton's laws in order to explain Kepler's laws. Methodologically, I can make that claim at least plausible. One remark —due to Alan Garfinkel—is that *a good explanation is invariant under small perturbations of the assumptions*. One problem with deducing Kepler's laws from the gravitational field equations is that if we do it, tomorrow the gravitational field equations are likely to be different. Whereas the explanation which consists in showing that whichever equation we have implies Newton's equation to a first approximation is invariant under even moderate perturbations, quite big perturbations, of the assumptions. One might say that every explanation of Kepler's laws "passes through" Newton's laws.

Let me come back to the philosophy of mind, however. If we assume a thorough atomic structure of matter, quantization and so forth, then, at first blush, it looks as if *continuities* cannot be relevant to our brain functioning. Mustn't it all be discrete? Physics says that the deepest level is discrete.

There are two problems with this argument. One is that there are continuities even in quantum mechanics, as well as discontinuities. But ignore that, suppose quantum mechanics were a thoroughly discrete theory.

The other problem is that if that were a good argument, it would be an argument against the utilizability of the model of air as a continuous liquid, which is the model on which aeroplane wings are constructed, at least if they are to fly at anything less than supersonic speeds. There are two points: one is that a

discontinuous structure, a discrete structure, can approximate a continuous structure. The discontinuities may be irrelevant, just as in the case of the peg and the board. The fact that the peg and the board are not continuous solids is irrelevant. One can say that the peg and the board only approximate perfectly rigid continuous solids. But if the error in the approximation is irrelevant to the level of description, so what? It is not just that discrete systems can approximate continuous systems; the fact is that the system may behave in the way it does *because* a continuous system would behave in such and such a way, and the system approximates a continuous system.

This is not a Newtonian world. Tough. Kepler's law comes out true because the sun-earth system approximates a Newtonian system. And the error in the approximation is quite irrelevant at that level.

This analogy is not perfect because physicists are interested in laws to which the error in the approximation is relevant. It seems to me that in the psychological case the analogy is even better, that continuous models (for example, Hull's model for rote learning which used a continuous potential) could perfectly well be correct, whatever the ultimate structure of the brain is. We cannot deduce that a digital model has to be the correct model from the fact that ultimately there are neurons. The brain may work the way it does because it approximates some system whose laws are best conceptualized in terms of continuous mathematics. What is more, the errors in that approximation may be irrelevant at the level of psychology.

What I have said about *continuity* goes as well for many other things. Let us come back to the question of the soul people and the brain people, and the isomorphism between the souls in one world and the brains in the other. One objection was, if there is a functional isomorphism between souls and brains, wouldn't the souls have to be rather simple? The answer is no. Because brains can be essentially infinitely complex. A system with as many degrees of freedom as the brain can imitate to within the accuracy relevant to psychological theory any structure one can hope to describe. It might be, so to speak, that the ultimate physics of the soul will be quite different from the ultimate physics of the brain, but that at the level we are interested in, the level of functional organization, the same description might go for both. And also that that description might be formally incompatible with the actual physics of the brain, in the way that the description of the air flowing around an aeroplane wing as a continuous incompressible liquid is *formally incompatible with the actual structure of the air*.

Let me close by saying that these examples support the idea that our substance, what we are made of, places almost no first order restrictions on our form. And that what we are really interested in, as Aristotle saw,[5] is form and not matter. *What is our intellectual form?* is the question, not what the matter is. And whatever our substance may be, soul-stuff, or matter or Swiss cheese, it is not going to place any interesting first order restrictions on the answer to this question. It may, of course, place interesting higher order restrictions. Small

effects may have to be explained in terms of the actual physics of the brain. But when we are not even at the level of an *idealized* description of the functional organization of the brain, to talk about the importance of small perturbations seems decidedly premature. My conclusion is that we have what we always wanted—an autonomous mental life. And we need no mysteries, no ghostly agents, no *élan vital* to have it.

NOTES

This paper was presented as a part of a Foerster symposium on "Computers and the Mind" at the University of California (Berkeley) in October, 1973. I am indebted to Alan Garfinkel for comments on earlier versions of this paper.

1 Even if it were not physically possible to realize human psychology in a creature made of anything but the usual protoplasm, DNA, etc., it would still not be correct to say that psychological states are identical with their physical realizations. For, as will be argued below, such an identification has no *explanatory* value *in psychology*. On this point, compare Fodor (1968).
2 Joke credit: Joseph Weizenbaum.
3 This is one of Kepler's Laws.
4 Provided that the two bodies—the sun and the earth—are the whole universe. If there are other forces, then, of course, Kepler's law cannot be *exactly* correct.
5 E.g. Aristotle says: 'we can wholly dismiss as unnecessary the question whether the soul and the body are one: it is as meaningless to ask whether the wax and the shape given to it by the stamp are one, or generally the matter of a thing and that of which it is the matter.' (See *De Anima,* 412 a6–b9.)

REFERENCE

Fodor, J. (1968) *Psychological Explanation,* New York: Random House.