Patrick Suppes

CAN PSYCHOLOGICAL SOFTWARE BE REDUCED TO PHYSIOLOGICAL HARDWARE?

The question of the title I answer in the negative. There are four strands to my argument. The first, which corresponds to the first section of the paper, analyzes the nature of computation. The second concerns the nature of goal-oriented behavior. The third uses an argument that the mind is computationally irreducible. The fourth asserts the irrelevancy of the standard attempts to provide a reduction via general ideas about determinism.

I. Nature of Computation

We may stipulate, I believe, for this paper that the mind is among other things a computational device. This means that matters of computation are of central importance in any arguments about the reduction of psychological concepts to physiological ones. Part of my argument about the irreducibility of computational concepts of the mind to physiological concepts is from the much simpler case of digital computers. In the case of digital computers, we understand to a very much more thorough degree exactly the physical basis of computation and at what level the computational concepts interface the physical concepts. In spite of this great precision of knowledge of interface, we do not at all attempt in the standard theory of computation for digital

E. Agazzi (ed.), The Problem of Reductionism in Science, 183–198.

computers to replace computational concepts by physical ones, which corresponds to replacing psychological concepts by physiological ones. In fact, I am skeptical that even in the case of relatively simple digital computers we could make *direct* physical observations on the computer from which we could infer in detail and without any high degree of error what from a software standpoint was actually being computed. The situation is very much more complicated and difficult, and less likely ever to be understood thoroughly, in the case of our own mental computations, because there is no evidence we will make much headway on the detailed physiological or physical identification of the neurons that are doing any particular computation. Notice of course that this lack of clear identification of physical location, quite apart from understanding the details of that physical location is characteristic of computation in a digital computer. It is a day dream to think that we can easily identify where a particular computation is taking place. Computations in a modern computer move around dynamically. Where they are even placed initially is not a static concept but a dynamic one depending upon what is present and what else is being computed at the time computation is started. It would be a great surprise if something similar is not true of the computations in the brain. Location of mental computations in terms of individual neurons seems totally out of reach. Global location of computations of a particular kind being done in a particular region of the brain is sometimes feasible.

It is also important that physically very different computers compute the same function even by different software programs. Only isomorphism at a high level is usually of interest, and really never in terms of concepts of computation at the level of individual transistors, or to be even more reductionist, at the level of individual elementary particles which make up the many different microscopic parts of a transistor.

Still another concept that seems likely to hold for the brain is that neurons compute statistically, unlike most of the current digital computers. With a statistical computation it is especially unfeasible to think about a reduction from the software to the

hardware. Each neuron is making a statistical contribution, but the physical performance of a particular neuron is not of any decisive importance. The feeling is rather here that we have something like the standard result for random variables in probability theory. Given the random variables, which roughly speaking are meant to correspond here to the software concepts, we only have requirements of consistency for there to exist a common sample space. The sample space is never unique. The basic consistency theorems, for example, Kolmogorov's theorem, deal with the existence of an underlying sample space. Without a very large number of assumptions that are not part of the standard theory, there is no unique sample space. The same is surely the situation with the statistical computation of neurons. The underlying neurons actually making the computation can in all likelihood never be observed, —certainly not with present concepts for observation of neurons. A given computation in the brain may be located in different neurons depending upon the dessert one had for lunch or the last green perception that passed through the system. There is such intensive contextualism that a reasonable conjecture is that identifiability for purposes of reduction is out of the question.

In order to eliminate fairy tales, it is important to tighten the argument here and to say that the formal claim of reduction not being possible is relative to a set of observational variables and observational techniques. I will not in this paper attempt such a formalization but I certainly think it is possible and can be done in a straightforward way for simple examples. Obviously, I am not intending to give an *a priori* argument that will hold for all time regardless of what scientific methods are in place a thousand years from now. I am concerned to give at a foundational level an argument in terms of current science and relevant philosophical concepts, an argument that is meant to be a strong one from a computational standpoint, against any possibility of reduction relative to any set of currently observable concepts.

There is one point that might seem elusive in making the distinction between software and hardware in the case of biological organisms as opposed to digital computers. Of course,

even in the case of computers, the distinction is not as sharp as it might seem, for in some sense the software program must become a part of the hardware, i.e., a part of the physical organization of the computer. Where, it might be asked, does hardware stop and software begin. Once the software is embodied in the computer, as it is in a different way in the brain, this is not an easy question to answer from purely physical considerations in the case of the computer or physiological ones in the case of the brain. We are able to make the separation in the case of digital computers only in terms of knowledge of what has been done in a deliberate fashion to program the computer, and how the computer hardware has been organized to embody programs. Even this formulation for digital computers is much too simple. First, as already pointed out, the physical location of particular pieces of the program in the memory and processors of the computer is a matter of dynamic allocation and not something we can physically easily directly describe. More importantly, with the current emphasis on digital computers acquiring the capability of learning, the details of the program will not necessarily always be possible to identify.

At least at present our main way of thinking about the brain's software is just in terms of the kind of mental or behavioral concepts psychologists have been developing for a long time and the language of common experience for a much longer time. Moreover, it is undoubtedly these concepts that are the significant targets of any reductionist program. There is no doubt a more detailed and extended sense of software that could be defined. We could include the structural and functional computational changes to be attributed to learning rather than genetic inheritance. Whichever view is taken, reduction to the brain's hardware has no present hope of being carried out in detail.

II. Goal-Oriented Behavior

A second strong argument against reduction of psychology to physiology is to be found in goal-oriented behavior in men and other animals. If I ask my well-trained dog to fetch the newspaper, which has just been delivered, no amount of physiological or physical observation of the dog would be able to either predict the trajectory of his motion as he goes in search of the paper, or infer what task he was attempting to carry out. I am not suggesting that the activity of the mind operates without use of the brain, it is just that psychological concepts cannot be reduced to physiological ones. In other words, we need psychological concepts and the theories that embody these concepts as theories that can be proved independent from a logical standpoint of purely physiological theories. It is a scientific fantasy to think we shall ever be able to make within our present scientific framework sufficient observations on my dog or on any other to determine what task it is engaged in, if those observations are restricted to purely physiological, including neurological, methods of observation and analysis.

To draw a drastic parallel —perhaps too drastic for some-, consider the use of logic in formulating physical theories. No one would be so foolish to say that we can reduce physics or physical concepts to a matter of logic or logical concepts, just because we need logic in the formulation of physical theories. This is how I see the relation of psychology and physiology.To talk about the brain as a machine that we can understand mechanically, which is familiar talk among physiologists and even neurologists, is to talk in a mistaken way. Think how absurd it would be to hear physicists talking about physics as a logical subject and meaning by that that only logic was needed to formulate physical theories. (It is as if logical methods for rigorously proving the independence of axioms or concepts had not been developed for the past hundred years.)

There is another point to be made about goal-oriented behavior. It is easy enough for almost everyone to accept the fact that from purely physical or physiological observations no one

can predict where I am going as I leave the house on a complicated physical trajectory to my office, to another house, to a store, or to a restaurant. However, it could be claimed and would be by some die-hard reductionists, that this is simply a case of unpredictable behavior, also to be found in purely physical systems, but unpredictable behavior that can be explained by purely physical concepts. The separation of explanation from prediction is an important matter scientifically but doesn't really bear on the present case, in the following sense. No matter how many observations or how much information of a purely physiological or physical sort is to be collected prior to or after I execute my chosen route to restaurant, store, or whatever, it will be impossible on purely physiological or physical grounds to explain the complicated path I followed. The man on the street recognizes such an enterprise as nonsense. It is important to recognize that from the most fundamental scientific standpoint it is nonsense as well to think in terms of being able to make such a purely physiological or physical analysis of my or any other higher organisms complicated movements.

Physiologists sometimes talk about goal-oriented behavior in cells. Without attempting to judge the scientific merit of this line of analysis, it is evident that we do not have the faintest idea of how to reduce the goal-oriented behavior of a higher organism to goal-oriented behavior of its individual cells. In other words, global goal-oriented behavior cannot be successfully analyzed in terms of goal-oriented behavior of cells. There is a seductive analogue here that could lead to mistaken conclusions. The analogue is that of the reduction of thermodynamics to statistical mechanics. In this case, the behavior of macroscopic parts of matter is reduced at a physical level for certain concepts to the behavior of microscopic particles. Moreover, the relation, as suggested above for neurons in mental computation, is statistical. However, the enormous difficulty of making this reduction a rigorous one, even under the simplest conditions, shows how improbable it is that at any time in the foreseeable future of science we would have the faintest idea of how to carry through a serious reduction of global goal-oriented behavior of an organism

to behavior of the organisms individual cells. There is something enormously seductive about this analogue. It is natural to think that we should somehow be able to push through a program of reducing our ordinary behavior as persons to the structure and function of the many billions of cells that make up our bodies. It is a metaphysical dogma that is hard to dislodge. My point is not to say that I can prove it as false, but just to say that the evidence for it is negligible. A way of putting the point is that it is very unlikely that the concepts needed to describe the global behavior of an organism can be reduced to concepts applicable only to individual cells.

It is important to block one mistaken conclusion from the argument I have just given. What I have said is not meant to suggest that we cannot identify at a level even below that of individual cells, for example, at the level of DNA or at the level of genes, microscopic features that predict major features of behavior. The triumphs of genetic analysis of many different sorts of diseases is a major triumph, and something very special about science in this century. It is on the other hand, idle to think that we have even begun to touch the problem of being able to infer goal-oriented movement of an organism from cellular observations. The usual scientific pluralism is at work here. We can do some things well, but not others, in terms of reducing global aspects of behavior to molecular ones. It is a form of metaphysical imperialism —in my view a scientific mistake— to think that we can generalize the successes of molecular biology to carrying out anything like a reduction of all major aspects of goal-oriented behavior. I mention again the difficulties that have been encountered in the last two decades in carrying through in a rigorous way the program of reduction of thermodynamical systems to statistical mechanical ones. It is easy to give from the current literature examples of thermodynamical systems that we do not know how to reduce to statistical mechanical systems. The incomparably more subtle and difficult problem of reduction of the theory of movements of higher organisms seems scientifically totally out of reach.

III. Computational Irreducibility

A familiar important, but not always remarked upon, property of classical physical systems that have been the object of much attention in the history of physics is the property of being computationally reducible. Here is the simplest and most important example. Newton's solution of the two-body problem, i.e., the problem of motion of two bodies acted upon only by the forces of their mutual gravitational attraction, permits us to predict the motion in the future or the past, or the position at any future or past time, given appropriate data on initial conditions at a given time. Moreover, this model has the important property of being applicable in first approximation to two-body systems for the planets, with the sun as one of the bodies. The fact that we can solve the equation of motion in closed form and thereby compute quite directly the future path of the bodies is of fundamental importance. Unfortunately, there was for a long time the feeling that this would be the norm for physical systems, i.e., that most of them were computationally reducible. We would be able to solve the equations of motions to determine the paths of the particles for any indefinite time into the future. However, already in the nineteenth century, the intractable problem of moving from two bodies to three bodies gave plenty of evidence that our ability to computationally reduce most physical systems was probably extraordinarily limited. Moreover, even when we cannot solve the equations of motion in closed form, we often feel that we can do a very good job of a numerical approximation. Already, however, in the case of the three-body problem, as was essentially shown by Poincaré, this was not possible. Now we understand the phenomena very thoroughly for systems that are drastically unstable, as is the situation for some initial conditions in the three-body problem. The numerical methods of computation, necessarily approximate in character, provide a very limited horizon of practical computation concerning the behavior of the system or, to put it in other terms, a very limited horizon of predictability for the behavior of the system. Many other systems

of a similar simple physical character have now been identified. With the modern intense interest in chaotic systems we have a sense of limitations in principle about predictability and about computational reducibility of physical systems that did not exist until rather recently, even in so well-established an arena as that of classical mechanics.

There are many reasons to think the mental computations of the mind are also computationally irreducible. One consequence of this is that we shall not be successful in simulating artificially the behavior of the brain. We shall not be successful in the sense that important aspects of human behavior will be missed in any such simulation. Even a model of ten billion artificial neurons will be deficient in providing anything like predictive or computational models. Notice that what we would like is really hopeless: speed up the computations of the brain by four or five orders of magnitude with a model of three or four orders of magnitude less neurons and thereby predict rather well by such computational reducibility future behavior. An unlikely story if ever there was one. To argue that the mind is computationally reducible, as in the case of arguments about physical systems, does not mean that we cannot find subsystems or aspects of behavior that can be computationally reduced. In other words, we can make certain theoretical computations about the behavior of the system that can be verified, and the computation is much simpler and faster than the behavior of the system itself. For example in the case of the three-body problem we can compute for restricted cases the escape velocity of one body, and we can make predictions about the behavior of a qualitative sort without being able to make computations about the detailed behavior. But just as in the case of such physical systems, the brain cannot be computationally simulated in simpler fashion, i.e., be reduced to the computations of the simpler system, when we are concerned with its full behavior. It is a piece of unrealizable scientific fantasy to think that we can move our minds to simulated brains and preserve our psychological identities. There is no reason whatsoever to think such a computational transfer will ever be possible. Above all, the physiologists and

neurologists will never make a computational reduction to formulas that lead from individual cell behavior to the mental computations of a fully functioning brain. I am not proposing to offer a metaphysical proof that such a neurological reduction is impossible, it is just that there is no serious scientific evidence whatsoever that it ever will be achievable within the framework of science as we now conceive it.

If I am right in this last claim, it means that in any serious formal or scientific sense reduction of psychological concepts to physiological ones will not be possible. Here what I state informally I mean in a more formal way. Given a formally and empirically adequate psychological theory of psychological phenomena, it will not be possible to prove a representation theorem in terms of a formally and empirically adequate physiological or neurological theory. A certain kind of handwaving may be indulged in by reductionist-minded philosophers, but no serious demonstration of reduction will be given, and it will not be given for substantial reasons. There is no scientific evidence that such a reduction can be carried out at a satisfactory level of detail.

There is still another and different point to be made about computational irreducibility. If we think of having a uniform theory of neurons —meaning that neurons act in the same way from one individual to another and their interaction with software is the same, reduction seems unfeasible. The hopelessness of the situation increases even more when we introduce the hypothesis, which seems likely to have considerable support, that the way in which individual neurons in a given individual interact with software is different from person to person. This would be a natural consequence of biological development occurring in a partially random fashion at the level of dendritic formation and the learning experience in terms of which some of that development is influenced also occurring with random variation from one individual to another. This would mean that the hardware of the neurons is connected to the software of thought in quite different ways in different individuals. If this conjectured variation from individual to individual holds, then reduction is all

the more impossible. The detailed structure of computations in one individual, taking the hardware and software together, would differ in quite significant ways from the corresponding structure in any other individual.

IV. Irrelevance of Physical Determinism

The argument for reduction of psychology to physiology, as a byproduct of the reduction of physiology to physics, as a consequence of determinism, is usually not put in as direct and simple a way as I will put it here. I think, however, the force of the argument is as follows. The physical universe is deterministic because as some analytic philosophers, untrained in physics, would put it, it is analytic that like events must have like causes. Given that there is a deterministic account of the physical universe, it then follows that everything that takes place in that physical universe is equally determined. If we know all there is to know about the physical world, that will fix uniquely all the other phenomena including, of course, the mental activities of higher organisms.

Philosophers have been ringing the changes on this argument with different degrees of explicitness and different degrees of emphasis for a long time, at least since the appearance of Kant's *Critique of Pure Reason* in the latter part of the eighteenth century. Kant doesn't discuss explicitly the concept of determinism, for it had not really surfaced in a completely clear way, even though there is a famous passage about the deterministic nature of the universe in LaPlace's introduction to his treatise on probability which appeared not long after the publication of Kant's *Critique*. But there is no doubt that Kant implicitly adopted a deterministic view in his using classical physics in generalized form as the metaphysical foundation of natural science and in his treatment of the category of causality in the *Critique of Pure Reason*.

Of course, Kant was not for a moment prepared to adopt the view that psychology could be reduced simpliciter to physical

determinism. He, as in his treatment of the antinomy of free will, was quite prepared to bite the bullet and remove psychology entirely from the domain of science, or as he would put it, more restrictively, natural science. The radical character of Kant's solution to the acceptance of physical determinism, as I would put it, has not always been properly recognized when it comes to working out what one could then hope to do with the science of psychology, but he certainly lays out his views in as explicit a way as could be asked for in the following passage in the preface to the *Metaphysical Foundations of Natural Science*:

> But the empirical doctrine of the soul must always remain yet even further removed than chemistry from the rank of what may be called a natural science proper. This is because mathematics is inapplicable to the phenomena of the internal sense and their laws, unless one might want to take into consideration merely the law of continuity in the flow of this sense's internal changes. But the extension of cognition so attained would bear much the same relation to the extension of cognition which mathematics provides for the doctrine of body, as the doctrine of the properties of the straight line bears to the whole of geometry. The reason for the limitation on this extension of cognition lies in the fact that the pure internal intuition in which the soul's phenomena are to be constructed is time, which has only one dimension. But not even as a systematic art of analysis or as an experimental doctrine can the empirical doctrine of the soul ever approach chemistry, because in it the manifold of internal observation is separated only by mere thought, but cannot be kept separate and be connected again at will; still less does another thinking subject submit to our investigations in such a way as to be conformable to our purposes, and even the observation itself alters and distorts the state of the object observed. It can,

> therefore, never become anything more than a historical (and as such, as much as possible) systematic natural doctrine of the internal sense, i.e., a natural description of the soul, but not a science of the soul, nor even a psychological experimental doctrine. This is the reason why in the title of this work, which, properly speaking, contains the principles of the doctrine of body, we have employed, in accordance with the usual practice, the general name of natural science; for this designation in the strict sense belongs to the doctrine of body alone and hence causes no ambiguity.[1]

Unfortunately, not many philosophers, and I would say almost no scientific psychologists, would be prepared today to follow Kant's way out of the dilemma of determinism.

Of course one immediate response to what I have called the dilemma of determinism is the modern one of saying that quantum mechanics has shown that the microscopic world of physics is not deterministic. I am not going to take that line of argument here because I don't believe it. My own view about quantum mechanics, expressed in several places, is that quantum mechanics is a weak probabilistic theory of the mean.[2] In this view, quantum mechanics is compatible with both deterministic and indeterministic hidden-variable theories, of which perhaps the best example of the latter is stochastic mechanics as developed by Edward Nelson and others, with the understanding that classical Markovian assumptions of Brownian motion must be relaxed to deal with problems of locality. But also as a viable possibility is the kind of deterministic hidden-variable theory outlined by David Bohm and various colleagues. I do not claim fully to understand Bohm's ideas and they are yet to be given an articulated and detailed development, but there is no reason to think that they cannot in principle be elaborated. The difficulty, of course, with any of these extensions of quantum mechanics, in the sense of providing a hidden-variable theory, —one that takes proper account of locality problems—, is being able to make an

experimental determination as to whether or not the theory is correct. Such theories may get defeated in the second round by their attempts to go beyond the phenomena of classical quantum mechanics to relativistic particle phenomena, quantum electrodynamics, and more generally to the wide range of experimental findings in elementary particle physics.

What I want to argue is that whoever is right about the proper hidden-variable theory for quantum mechanics, which may turn out to be a purely metaphysical choice, determinism as a general thesis is irrelevant to the question of the reducibility of psychology to physiology. The reason for my holding this view is easy to state. Determinism is too capacious and general a theory to help any such issue to be settled in an interesting way. Why is this? Because the collection of theories that are deterministic is able to accommodate any sort of behavior. Perhaps the way to illustrate this without too many complications and reservations is to consider again the physically simple case of the three-body problem that I have discussed elsewhere with reference to propensity theories of probability.[3] As we move from two-bodies to three-bodies, our detailed understanding of the motions of the three bodies disappears, a fact which I pointed out earlier has been well-known since the nineteenth century. What has not been well-known since the nineteenth century is the proof that for rather simple restricted cases of the three-body problem —meaning a reduction of the problem to the motion of a single-body where the motion of that body is determined by the other two bodies— the following sorts of results hold. First, there exist initial conditions, which in this case are just the initial position in one-dimension and the velocity in one-dimension of the body, such that the sequence of the largest integer values contained in the temporal half-cycles of passing through the plane of the other two bodies has the following property. The sequence of integers so generated, the so-called symbolic dynamics, can represent any random sequence of integers, where the integers are greater than a certain constant. We can therefore represent in terms of the symbolic dynamics of this simple deterministic system -simple in terms of understanding its causes and the derivation of the

differential equation governing its motion, not simple in terms of its actual motion-, any random sequence of heads and tails. Second, in contrast there exist initial conditions, in this case just a single number, the velocity in one-dimension of the body, such that the symbolic dynamics encodes the contents of the books in the Library of Congress. One example is purely random, the other is as intentional as you wish, but the richness of this simple, deterministic system is capable of generating either phenomena.

As I have argued in another paper[4] we can develop the same line of attack with purely indeterministic systems. If you don't like determinism choose indeterminism. This choice corresponds to the two choices of hidden-variable theories for quantum mechanics I mentioned above. So does one choose between indeterministic structures of some general probabilistic theory or unstable structures of some deterministic theory? Put in very broad terms the choice seems to be a matter of taste in metaphysics. At the present stage of science there seems no likelihood of any sequence of crucial experiments that will force one of the two positions to the wall. Indeterminism and determinism are here to stay. Exercise your metaphysical choice as you will. There is no inconsistency between determinism and randomness. We can use unstable deterministic systems to generate any probabilistic phenomena desired, or we can take a system that is indeterministic and not known to be deterministic, if that is your metaphysical bent. There are, in fact, some beautiful theorems by Donald Ornstein and his colleagues that make the metaphysical point in still stronger fashion: there are physical systems on which we can make an infinite number of observations —or if you want more precision, there are mathematical models of certain physical phenomena—, such that on the basis of these infinite number of observations it is impossible to distinguish between a deterministic mechanical model governing the phenomena and a stochastic process governing them. This line of argument, the last line of argument I consider here, can be summed up in this way. The classical attempts to reduce psychology or our mental concepts to physical theories and physical concepts by general deterministic arguments

is to try to reduce the rich facts of our mental activity to an unverifiable metaphysics of determinism. Kant had the story upside down: our mental life is empirically a rich phenomena which we can study scientifically and successfully. In contrast the general theory of determinism as a view of the universe represents a metaphysics empty of content.

Stanford University (USA)

NOTES

1. *Metaphysical Foundations of Natural Science*, translated by James Ellington, Bobbs-Merrill Company, Incorporated, New York, 1970, pp. 8-9.

2. Probabilistic Causality in Quantum Mechanics. *Journal of Statistical Planning and Inference*, 1990, 25, 293-302.

3. Propensity representations of probability. *Erkenntnis*, 1987, 26, 335-358.

4. Indeterminism or Instability, Does it Matter? In G. G. Brittan, Jr. (Ed.), *Causality, Method, and Modality*, Kluwer Academic Publishers, 1991. Pp. 5-22.