The Journal of Mind and Behavior
Winter 1989, Volume 10, Number 1
Pages 1–20
ISSN 0271-0137



# Consciousness and the Incompleteness of the Physical Explanation of Behavior

Avshalom C. Elitzur

*Weizmann Institute of Science*

All theories attempting to solve the mind-body problem on the basis of a scientific world-view assume that the present framework of physics is, in principle, a sufficient basis for a complete causal explanation of behavior. From this assumption follows another assumption, namely, that consciousness is a completely passive phenomenon, devoid of any causal role. These two assumptions were hitherto considered impossible to prove or disprove. In this article I review the theories based on the passivity assumption and stress their rationales. I claim, however, that there is one instance in which behavior cannot be explained without assigning a causal role to consciousness. The argument is then generalized: a causal influence of consciousness, in addition to all known physical forces, is claimed to underlie other modes of behavior as well. Possible objections to this argument are encountered and answered. The issue is further discussed in connection with evolution, artificial intelligence and thermodynamics.

This paper presents an unorthodox argument, namely, that it is impossible *in principle*, even in a theoretical idealized case, to give a complete causal explanation of human behavior on the basis of the present system of natural laws. This claim is based on discussions that may appear more philosophical than scientific. This, however, is only due to the enormous complexity of the phenomena under study. The argument itself, if valid, has a direct bearing on physics' most basic assumptions, and on some fundamental issues in modern science in general.

## *Does Consciousness Constitute a Real Problem for Science?*

At the end of *Mind and Matter* Schrödinger (1958) summarizes a paradox that retained its acuity since the days of the Greeks:

I wish to express my gratitude to Prof. J. Rosen, Prof. L.P. Horwitz, and Prof. B.-A. Scharfstein, of Tel Aviv University, for several enlightening discussions and valuable comments. Thanks to Prof. H. Putnam of Harvard University for a heated yet delightful discussion and for drawing my attention to some relevant sources. Prof. A. Shimony of Boston University proved to be a very sensitive and stimulating critic of this paper. Jessica Regan and Ruth Goldstein kept an eye on my English. Last but not least, I am grateful to Prof. Schneior Lifson and the Department of Chemical Physics, Weizmann Institute of Science, Rehovot, for their kind hospitality during the period this paper was written. Requests for reprints should be sent to Avshalom Elitzur, Department of Chemical Physics, Weizmann Institute of Science, 76 100 Rehovot, Israel.

> In this chapter I have tried by simple examples, taken from the humblest of sciences, namely physics, to contrast the two general facts (a) that all scientific knowledge is based on sense perceptions, and (b) that none the less the scientific views of natural processes formed in this way lack all sensual qualities and therefore cannot account for the latter. (p. 103)

On the other hand, Niels Bohr, fascinated by the notion of complementarity, believed that this notion could reconcile all the great debates in biology and psychology (e.g., determinism vs. teleology) simply by accepting both opposing views as complementary. Hence he saw the incompatibility of the physical and the mental as obligatory rather than puzzling.

> . . . For describing our mental activity, we require, on the one hand, an objectively given content to be placed in opposition to a perceiving subject, while, on the other hand, as is already implied by such an assertion, no sharp separation between object and subject can be maintained, since the perceiving subject also belongs to our mental content. From these circumstances follows not only the relative meaning of every concept, or rather of every word, the meaning depending upon our arbitrary choice of viewpoint, but also that we must, in general, be prepared to accept the fact that a complete elucidation of one and the same object may require diverse points of view which defy a unique description. . . . The necessity of taking recourse to a complementary, or reciprocal, mode of description is perhaps most familiar to us from psychological problems. (Folse, 1985, p. 179)

The problem that intrigued these authors is known as the mind-body problem or the problem of consciousness (Fodor, 1981; Popper and Eccles, 1977). Although the term "consciousness" is often used in some physical theories, almost nowhere in the physical literature has the fundamental problem associated with it been properly described. A few remarks will suffice here as an introductory presentation.

Physics aspires to be the most fundamental science, on the basis of which other sciences can be based. Hence all phenomena should, in principle, be explicable in physical terms. Even an ardent opponent to such a view cannot deny that complex phenomena such as brain mechanisms can be explained to a considerable degree by breaking them down into basic chemical and electrical events. Consciousness, on the contrary, notoriously evades any such an explanation. It is possible in principle to give a complete physical description of human behavior, yet nothing in such an account would indicate the existence of conscious experiences. This contrasts with other characteristics of physical phenomena. For example, a chemist confronting the formula $C_6H_5COOH$ for the first time will be able to recognize the chemical as an acid, because its acidic properties follow logically from its atomic structure. Even if no theory has yet been proposed linking the chemical's formula with its properties, the relation itself is possible in principle. In contrast, an ideal and complete description of the neurophysiological processes occurring between the moment of being pinched by a pin and the moment of uttering a cry of pain does not give any indication of a *subjective experience* of pain.

Neither does the existence of consciousness seem to be necessary in order to explain this behavior. As far as physics is concerned, this subjective feeling might just as well not occur at all. The causal chain of physical events in the brain suffices.

Ever since its beginning, science exhibited two primary approaches concerning the mind-body problem. One approach sought for something explicitly non-physical at the root of consciousness, thus tackling the problem at the cost of creating additional problems. The other approach was a more practical one; it restricted itself to the question of whether consciousness has any observable effect. Is the existence of consciousness necessary to explain behavior, or is it possible in principle to view behavior as a purely physical mechanism – however complicated – of stimuli and responses? If the second possibility is true, then consciousness should not bother the scientist, although it may remain an interesting occupation for the philosopher.[1] This way of putting the question – namely, can any effect of consciousness be observed – seems to be the most meaningful form of the mind-body problem posed to science. It will be posed in the following pages.

### *Consciousness Is Not Essential in Quantum-Mechanical Observation*

Let us begin with quantum mechanics, since it is the only field in modern physics where consciousness has been given considerable attention. It is disappointing to realize, upon a closer examination, that this path leads to a dead end.

The reasoning that led to the hypothesis concerning a unique role of consciousness in quantum mechanics is indeed an appealing one: Since measurement is an intriguing problem in quantum mechanics, and since consciousness is a mystery, and since all physicists are conscious beings, could the two problems be one and the same? This was Wigner's (1970, 1975) bold suggestion: he ascribed to mind the ability to "collapse" the wave-function non-locally (and, as we shall see, non-temporally). This model was elaborated by Eccles (1953, 1986), who sought a way in which free will might interfere with the microscopic neural discharges of the brain. Stapp (1982) proposed another detailed model describing both quantum-mechanical and neurological aspects of such a mechanism. The alleged active role of consciousness in observation became especially attractive for physicists who accepted parapsychological phenomena and tried to account for them by means of a new physical theory (e.g., de Beauregard, 1980; Jahn and Dunne, 1986; Schmidt, 1982).

Mainstream quantum mechanics, however, did not adopt this line of thought, and for good reasons. Several authors, adhering to a wide variety

[1]The status of consciousness in this view resembles that of the "hidden variables" in quantum physics, aimed to explain the known phenomena yet producing no testable prediction.

of schools (e.g., Bohm, 1951; Penrose, 1987; Popper, 1982; Wasserman, 1983; Wheeler, 1980, 1981), claimed that "measurement" could as well be carried out by a mechanical instrument, yielding similar effects. "A measurement," stressed Peres (1986a), "is not a supernatural event. It is a physical process, involving ordinary matter, and whatever happens ought to be explained by the ordinary physical laws" (p. 688). The difficulty in testing the hypothesis concerning the influence of consciousness lies in the fact that, according to the theory, the influence of consciousness on the observed system can be exerted even at a later time. The instrument may thus remain in a superposition, together with the event measured, until someone reads it. Now, since all physicists possess consciousness, this hypothesis lies entirely beyond proof or disproof and hence is metaphysical rather than scientific. Nor can parapsychology help to persuade the conservative critic, in view of its admitted failure so far to demonstrate a repeatable experiment (Jahn and Dunne, 1986).

As for free will, critics of Eccles' model argued that a neural synapse, although very small, is still millions of times larger than the sub-atomic particles governed by quantum laws (Wilson, 1976). In any case, the whimsical quantum indeterminism does not seem closer to free will than strict determinism. Peres (1986b), discussing the role of free will, argued that this notion is merely a special case of the strict determinism governing macroscopic physics, although much more complicated.

Quantum mechanics, to summarize, does not face a real necessity to take consciousness into account. Does this hold also for macroscopic physics?

### *The Passivity of Consciousness in Physical Theories*

The study of consciousness in relation to the ordinary physical world of living organisms has always been the realm of philosophers, although a few physicists (e.g., Penrose, 1987; Schrödinger, 1958; Stapp, 1982) contributed to it as well. And as 20th-century philosophy became more committed to scientific principles, the theories concerning this issue made increasing efforts to remain consistent with physics; in the common parlance, they became physicalistic.

In this respect, the most elegant solution of the mind-body problem is provided by these theories which maintain that consciousness is causally passive. These theories do not deny that consciousness exists, as deceased behaviorism used to do, yet they assure us that we can pretend it does not. Consciousness, so runs the argument, is a mere reflection of the physical processes in the brain, another aspect of them. This was also Bohr's opinion in the passage quoted above. Therefore consciousness *per se*, as distinct from the brain processes, does not constitute a real part in the dynamics of the neural mechanisms. In order to understand this conclusion, consider the movements

of a billiard ball. Do we need consciousness to explain its movements? No, because mechanics completely suffice. Now, consider photosynthesis, blood-circulation, the operation of the hypothalamus, the unconscious information-processing of the cortex, and, finally, conscious thought:[2] a simple extrapolation from basic physical interactions to more complex, biological and psychological ones, leads to the expectation that a complete explanation will one day be possible concerning thought and emotion (Chase, 1979; Rensch, 1970). Consciousness will then be a mere side-aspect of the brain processes (Wassermann, 1983). In this framework, any observable behavior that seems to be the result of consciousness can as well be explained on the grounds of brain-physiology alone.

An appealing theory along these lines is known as the "identity theory," according to which mental states and brain states are *the same*, though for some reason we perceive their subjective aspect as distinct from their physical one. This theory gave rise to several philosophical and semantic discussions, that in the course of the years became very complicated—and no less boring. It is not necessary, however, to get here into the debate, because identity theory too is committed to the passivity assumption that this paper challenges, thus challenging identity theory too. Notice how it follows from identity theory that consciousness is passive. When a person utters a cry of pain, the common-sense explanation is that he did it because he felt pain. Identity theory, on the other hand, employs another description: before the man uttered the cry, a certain process occurred in his brain, strictly obeying physical laws, and *this* process was the cause of the uttering of the cry. The man himself perceived this brain process as pain. This very reasonable explanation obliges consciousness to be passive: without denying the existence of conscious feelings, it argues that each such a feeling is a reflection of a strictly physical brain-process. Hence, physics does not need to take consciousness into account when explaining behavior.

This passivity assumption became widely popular. In fact, I think one may talk here about the dogma of passivity. As the neurophysiologist Sir Charles Sherrington (quoted in Schrödinger, 1958, p. 43) put it, "Physical science . . . faces us with the impasse that mind *per se* cannot play a piano—mind *per se* cannot move a finger of a hand." As early as 1928 Sir Arthur Eddington referred to the materialist who "regards consciousness as something which unfortunately has to be admitted but which it is scarcely polite to mention" (p. 348). For him, Eddington stated,

> We have associated consciousness with a background untouched in the physical survey of the world and have given the physicist a domain where he can go round in cycles without ever encountering anything to bring a blush to his cheek. (p. 348)

[2]Recall in this context the modern experiments, reviewed by Dixon (1981), showing that even complex ideation can occur unconsciously.

This approach flourishes in modern philosophy. For example, four recent papers (see Churchland, 1985; Levin, 1986; McMullen, 1985; Tye, 1986) have arrived at the same conclusion concerning "qualia," i.e., subjective experiences. To take the example used by these authors, an imaginary future scientist who is congenitally blind, yet has sufficient knowledge about electromagnetism and brain-physiology, would know as much about colors as we do. Nothing in our subjective experience of colors can add to his knowledge. Even Skillen (1984), having ridiculed the "bed-time story" that the psychologist and the physicist can do their jobs without disturbing each other, admitted being himself attracted to the view that physiological and mental processes are somehow identical. Wilkes (1984), who reached a similar conclusion concerning qualia (her favorite thought-experiment employing a congenitally deaf scientist who knew about sounds better than she did), posed the question in its most fundamental form in the title of her paper: "Is Consciousness Important?"—her answer being a decisive "No."

Such a denial of any causal role to consciousness is the gist of the theories known as epiphenomenalism (physical events can influence mental events but not vice versa); parallelism (physical events and mental events run parallel, governed by independent causal laws, and never affecting one another); identity theory or double-aspect theory (brain processes and mental states are two aspects of the same phenomena); and a variety of versions of these theories. Such solutions of the mind-body problem, it should be remembered, were laboriously constructed in order to preserve the present framework of physics, avoiding both the extremities of the mentalist "no matter" and the materialist "never mind."

*Is Consciousness Indeed Passive?*

Yet, the theories based on the passivity assumption leave the core of the problem out of the reach of any solution. They banish consciousness from the world of natural sciences to the tiny private domain of one's inner world. We are led, as Schrödinger (1958) bluntly pointed out, to an absurd situation: the most essential aspect of one's life is irrelevant from the physical point of view. Let us briefly review the attempts made so far to challenge the passivity dogma.

Several biologists found it hard to believe that consciousness, apparently a characteristic of higher organisms, could have developed as a mere side effect of the brain's evolution. Indeed, it is more likely that consciousness makes an observable difference, i.e., it affects behavior in a way that helps survival, and is thus a part of the evolutionary process. No one, however, succeeded in refuting the passivity dogma, which seems as strong in this case as in any other. The zoologist Griffin, in his book *The Question of Animal Awareness*

(1981), made strenuous efforts to find one clear instance of animal behavior that proves the operation of consciousness. He shared with several other scientists the commonsense belief that animals possess some rudimentary degree of consciousness (e.g., Dobzhansky, 1967; Rensch, 1970; Schrödinger, 1958). Yet the result, as reviewed by Chase (1979) and Darden (1983), and as Griffin himself admits, is not impressive. All the instances of behavior he studied can as well be explained without assuming the involvement of consciousness (Wilkes, 1984).[3] Many authors (e.g., Chase, 1979; Rensch, 1970) went as far as to argue that evolution and human history would proceed exactly the same way, even if all living creatures lacked consciousness. Chase (1979) and Darden (1983) correctly point out that Griffin failed to prove the existence of consciousness in animals because he did not face a more fundamental problem: we cannot prove the existence of consciousness even in other human beings, though intuitively we are sure of it (Fodor, 1981; Schrödinger, 1964). The question of animal awareness is therefore only a particular case of the old problem of other minds.

Wilson (1976) and Culbertson (1977) attempted to object to the passivity dogma by means of what may be called the cogito argument: Since people claim to have consciousness, does this not indicate an active influence of their consciousness upon their physical behavior? As Dewan (1976) has shown, it does not. The use of the word "consciousness" may as well reflect mere learning of the human vocabulary, and thus cannot indicate more than brain-processes.

Natsoulas (1988), at the end of a long paper in which he praised behaviorism, asserted that "when we report the experience that we are now undergoing, *we must be directly (reflectively) aware of the experience, and choose our words accordingly*" (italics original). He did not try to give any support for this claim but concluded with an over-confident statement: "There is a good chance that radical behaviorists will disagree with my analysis. However, I must conclude: You can't leave it to the physiologist" (p. 54). One can sympathize with Natsoulas' wish to end the debate by whatever strong statement, but this is far from a scientific reasoning.

Shoemaker (1975) tried his luck by putting the cogito argument in terms of the qualia problem.[4] He reasoned that qualia are the cause of people's belief

[3]The role of consciousness in evolution has been questioned also by Mott (1982), although in a peculiar way. On the basis of a previous definition of "function," he concluded that i) function is exclusive, therefore ii) no two systems can fulfill the same function, hence iii) if consciousness has a function, then the corresponding brain-processes have not. The awkwardness of this reasoning already becomes conspicuous in (ii), since anyone who knows something about animals knows that there *are* several cases in which more than one organ fulfills a certain function. But Mott, rather than going back to see what was wrong with the definitions that led to this statement, went on to conclude that consciousness cannot have a function.

[4]See the above discussion of the deaf/blind scientist.

in qualia. Similarly Cuda (1985), during one of the science-fiction games that philosophers of mind are so fond of, reasoned about an imaginary Fred whose brain is replaced by an analog simulator, that if this unfortunate fellow "loses his ability to have red qualia, then he also looses his ability to believe he had red qualia, his ability to desire to have or not have red qualia, etc." Oddly enough, both Shoemaker and Cuda disclaim any far-reaching metaphysical ambition in this argument, which, if pursued systematically, poses a serious challenge to all physicalistic theories. Anyway, even Shoemaker's more modest claim can be rejected by the above counterarguments against the cogito argument. White (1985), replying to Shoemaker, concluded that "if one's belief that one's states have experienced features is caused in part by their having them, it is not caused by one's *experiencing* that that is so" (p. 380). I do not pretend to have completely understood this sentence, but White seems to state that one's belief in having qualia is not caused by these qualia.

To summarize, those who hold that consciousness is passive dismiss the cogito argument as a mere linguistic problem, and the dismissal is logically valid. Whether one says "I have pains," "I have hopes," or, ultimately, "I have consciousness," these statements may be caused only by physical processes in one's brain. Language cannot transfer conscious qualities. Any verbal "output" may be viewed as a mere result of the learned "input." Putnam (1960/1964), discussing artificial intelligence, stressed that

> . . . machine performances may be wholly *analogous* to language, so much so that the whole of linguistic theory can be applied to them. If the reader wishes to check this, he may go through a work like Chomsky's *Semantic Structures* carefully, and note that *at no place is the assumption employed that the corpus of utterances studied by the linguist was produced by a conscious organism.* (p. 95; italics original)

## *A New Argument for a Causal Role of Consciousness*

So, can we suggest an alternative hypothesis to the passive-consciousness dogma? Let us give it a try.

Perhaps, although any behavior seems to be explicable by physiology alone, this is so because consciousness is just a weak factor "outshined" by the neurophysiological processes, as a star's light is outshined by daylight. Consciousness, our hypothesis may continue, is an additional factor, although slight, due to which the mother's loving embance of her infant is *stronger* than that of a consciousness-lacking mother functioning due to brain mechanism alone; a rabbit escapes from the fox *faster* because it is afraid in addition to being conditioned to escape; a man with a toothache groans *louder* because of the additional conscious feeling of pain, and so on.

This hypothesis sounds hard to prove. It would seem that there is no instance of behavior that can be ascribed to consciousness that cannot be bet-

ter accounted for by physical mechanisms. Bearing in mind Griffin's failure (see above), we need to find an instance in which the explanation of a certain observable behavior would be clearly inadequate on the basis of neurophysiology alone, thus requiring the assumption that consciousness *per se* affects behavior.

I believe I am able to point out such a case.

What is the behavior for which only consciousness can account? Well, *consciousness must be the reason why people are bothered by problems of consciousness.* If someone says that he cannot understand his experiences by what he knows about himself, this expression of bewilderment cannot be explained by any physical process, unless one resorts to the far-fetched claim that the person expressing this bewilderment is lying. Otherwise, this negative statement, unlike the positive "cogito" statements, cannot be explained without assuming that consciousness, as a non-physical entity, affects speech, hence observable behavior and hence the physical world in general.

*Possible Objections*

Of course, the proponent of the passive-consciousness theories will not let us get away so easily. Anticipating his or her possible objections, we shall enter into a more complicated discussion, but in so doing we may push the opposing view into a narrower corner.

The objection to our argument might be put as follows. We seem to take for granted the validity of what people say. But what about the beliefs in ghosts, the evil eye and so forth? The "awareness" of a mind-body problem may be as illusory as any of these deeply rooted intuitions.

This objection, often made by "identity theory," is nothing but a new version of an old attempt to dismiss consciousness itself as an illusion. This attempt is invalid, as anyone who is not a radical behaviorist will agree. For the term "illusion" itself denotes two entirely different things: an information-processing malfunction that is purely physical, and a subjective conscious experience. As Popper and Eccles (1977, p. 208) have shown, a computer can also have illusions in the former sense without having consciousness. The mind-body duality is not dispensed with by invoking such terms, since these terms themselves suffer from the same ambiguity. Consciousness is not an illusion.

So this objection does not deny that consciousness exists, but only that it can affect our observable behavior. It thus boils down to the following standpoint: i) people have consciousness; ii) people express awareness of the problem of consciousness; iii) people express awareness of the problem of consciousness for reasons other than having consciousness.

But what are those other alleged reasons because of which people say there

is something mysterious about consciousness? Any attempt to reply while avoiding the simplest answer for which we opt must invoke a "will to believe," deliberate lie, or again, a new form of the illusion argument: the opponent must claim that *there are some physical processes in the brain leading to an illusion of consciousness*, in addition to the genuine existence of consciousness, and that this illusion only is the reason for expressing concern over the genuine problem of consciousness.

Quite awkward. Perhaps a thorough analysis could invalidate these arguments too, but I suggest that we leave them where they are and content ourselves with the following two achievements: first, the theories holding the passivity dogma have been reduced to an *ad absurdum* position that, I believe, had not been realized before. Second, these theories now seem to be committed to a *testable* hypothesis. If they are right, there must be some definite physical structure in the human brain responsible for the belief that consciousness evades physical description. The burdon of proof, here shown to be possible in principle, now rests on the opposing side.

### *Penrose's Argument and Its Clash with Physicalism*

The thesis of this paper is twofold. The first thesis is that consciousness is not passive but rather a part of the causation of behavior. The second and consequent thesis is that physics, being unable to describe consciousness, is inherently incomplete. About the time the first draft of this paper was written, a short paper by the notable physicist Penrose (1987) appeared, anticipating the first thesis. Penrose, however, did not pursue his argument to its far-reaching conclusion, i.e., the second thesis about the incompleteness of physics. The following discussion may therefore help to clarify how the latter thesis stems from the former.

First, here are Penrose's words.

> I would contend that the evolutionary development . . . of the ability to think consciously indicates that consciousness is playing an *active* role and has provided an evolutionary advantage to those possessing it. For various reasons I find it hard to believe that conscious awareness is merely a concomitant of sufficiently complex modes of thinking—and it seems to me clear that consciousness is itself *functional*. . . . Indeed, if consciousness had no operational effect on behavior, then conscious beings would never voice their puzzlement about the conscious state and would behave just like unconscious mechanisms "untroubled" by such irrelevancies! (p. 116; italics original)

Penrose challenges, as we do, the passivity dogma. Let us recall that the proponents of this dogma aimed to rescue the completeness of the physical worldview, because they admitted that consciousness is a *non-physical phenomenon*. Consequently, if one believes that consciousness can affect behavior, one thereby admits a serious gap in any physical explanation of behavior.

Does Penrose have such a heresy in mind, or does he adhere to a certain identity-theory version whereby consciousness is only another name for some physical process?[5] Recalling other unorthodox ideas in Penrose's work, it seems to me that he would not object to the more radical conclusion derived from the rejection of the passivity dogma in his paper. "Any world in which minds can exist," he says at the end of his paper, "must be organized on principles far more subtle and beautifully controlled than those even of the magnificent physical laws that have been so far uncovered" (p. 118).

*Consciousness and Artificial Intelligence*

Perhaps nowhere is the problem of the causal role of consciousness more pertinent than in the widespread debate about conscious machines. Can our new argument be applied to this issue?

The simplest approach in the context of machine-consciousness is that of positivism: "Only what we can observe matters." And since positivism made such a tremendous impact on physics at the beginning of the century, it is not surprising that Bohr applied this approach to the problem of consciousness. In a lucid exposition of Bohr's view Wheeler (1981) advocated an operationalist definition of consciousness, and claimed that there is no critical experiment that can do better. He suggested that once an intelligent computer can refer to itself, it fulfills the requirement for the definition of a conscious being. Wheeler quoted Bohr's further statement: "The question whether the machine *really* feels or ponders, or whether it merely looks as though it did, is of course absolutely meaningless" (p. 94).

The use of such overconfident words as "of course" and "absolutely" looks like an attempt to suppress an intriguing riddle. For, as our immediate experience tells us every moment, the fact that we really feel or ponder, and not merely look as if we do, is far from being meaningless. Here we see an example of the ambivalent nature of positivism: while it was right in seeing the assumption of non-local influence of consciousness as unnecessary in quantum mechanics, it is hard to follow it when it attempted to render the existence of consciousness meaningless altogether.

Yet, even today, no argument that could prove Bohr's claim wrong is known. Thagard (1986), while far from being a positivist, ends his discussion with an essentially similar conclusion: from the point of view of artificial intelligence, consciousness does not seem to provide any function.

Can our new argument suggest something better? If it is valid, then we have a better test to determine whether an intelligent computer is conscious. And

[5]Notice again the catch: if consciousness is only an aspect of a physical process, then consciousness is again passive—it is the *physical process* that influences other physical processes and consciousness can again be ignored!

perhaps—who knows?—this Gedanken experiment will be possible in the future. Let one ask that computer what feelings are. Will it reply "my feelings are this and that current in my wires," or will it report, bewildered, that there is something else about its feelings that escapes such definitions? A computer, it should be remembered, can know practically everything about its software via its designers. Likewise, due to our complete knowledge of its software, we can rule out the counterhypothesis of a deliberate lie.

But suppose that the definition of consciousness, as fed into our computer, were to be extremely precise, following the *Webster Dictionary* (1981), in which one of the definitions of consciousness is "something in nature that is distinguished from the physical." Would the fact that this definition was introduced to the machine from the beginning now cast doubt on its answer? No, because in this case the machine must be able to state whether it has or has not such a property that escapes physical explanations.

One may still argue that the computer must express a certain confusion concerning itself due to the system's inability to give a full self-description, as required by Gödel's theorem (Rucker, 1982). This claim can be put in the form of a more daring one, namely, that the whole mind-body problem reflects something similar to the inherent incompleteness of any self-description. Globus (1976) indeed suggested treating the mind-body problem along these lines. This analogy, however, is superficial. The unprovability of a certain logical statement, which can be made provable by introducing another unprovable one (Rucker, 1982), is obviously a different problem from the fundamental inexplicability of consciousness, that—and here lies its very epistemological peculiarity!—cannot be solved even *in principle* by any additional information.

Bohr was not alone in his adoption of positivism in the mind-body issue. In a celebrated paper, Turing (1950/1964) stated that a machine should be regarded as conscious if it can answer questions in such a way that the output cannot be distinguished from that of a human being answering those questions by the same mode of communication. This may still be reconciled with our criterion, once we refine it by requiring the computer to indicate confusion concerning consciousness. However, Putnam (1960, 1975) poses here a new challenge. He claims that any intelligent computer must raise questions similar to those raised by humans as to mind and body. When referring to itself, an intelligent machine will observe a difference between two modes of describing its own internal states: the engineer's structural blueprint and the logician's "machine table." These, claims Putnam, are analogous to the two possible descriptions of human psychology (1960, p. 84). Putnam even imagines a "mechanical Russell" that, like its human namesake, would be intrigued by the fact that experiencing something and having a certain internal state are not the same thing. He thus straightforwardly equates the mind-

body dichotomy with the software-hardware dichotomy and, further, with the linguistic "dichotomies" concerning water-$H_2O$ and light-electromagnetism (1960, p. 92). In a similar vein, Sperry (1969, 1980; see also a discussion in Stapp, 1982) likens the relation between consciousness and brain to the way in which "the properties of the molecule transcend the properties of its atomic components" (1969, p. 533). This echoes Putnam's recipe of rendering the mind-body problem a pseudoproblem, a tendency that Sperry himself warns against. Putnam's conclusion is straightforward and blunt:

> The moral, I believe, is quite clear: it is no longer possible to believe that the mind-body problem is a genuine theoretical problem, or that a "solution" to it would shed the slightest light on the world in which we live. (1960, p. 96)

Here again, with consciousness rendered an epiphenomenon, we seem to be pushed into the straight-jacket from which we wished to escape, where our problem has no right to be raised at all. If a machine must express similar questions concerning its perception of itself, then we cannot maintain that interest in the mind-body problem is proof of the effect of consciousness on the physical world.

A closer inspection of Putnam's arguments, however, reveals the awkward fact that, while dealing extensively with semantics, he managed not to talk about consciousness at all! The mind-body problem has nothing to do with the "problem" of the difference between two statements about a certain event. The latter problem is indeed merely linguistic. The mind-body dichotomy, on the other hand, splits each part of such an artificial dichotomy.[6] For example, I know that water is $H_2O$. The knowledge of this formula is indeed abstract, unlike the taste of water in my conscious experience. Yet the knowledge of the formula also has for me a conscious side: whether it is encoded in my memory as a visual or auditory representation of the formula, it always has a conscious aspect. It is *this* dichotomy, splitting both components of the above pseudo-dichotomy, that Putnam's computer fails to notice, thereby failing to meet our new criterion. In essence, Putnam's argument is merely another version of the "illusion argument" refuted above.

The fundamental mind-body problem will therefore never occur to a creature lacking consciousness, which can only perceive difference between hardware and software. This was pointed out sharply by Rucker's (1982, p. 183) penetrating distinction between three aspects of mind: hardware, software, *and* consciousness. In Putnam's argument we observe a common attempt to reduce a riddle concerning an essence to a mere problem of relations. Yet the core of the problem is not the gap between structural knowledge

[6]This was beautifully expressed by Schrödinger in the passage quoted at the beginning of this paper.

and consciousness, but the very existence of consciousness itself, a phenomenon not entailed—let alone explained—by anything we know.[7]

To conclude, Putnam's mechanical philosopher may be persuaded that the mind-body problem is not worth pursuing after reading its originator's paper (thus from a "mechanical Russell" turning into a "mechanical Putnam"), but this would rather distinguish it from a conscious and truly sensitive thinker. The question of whether an artificial-intelligence machine possesses consciousness can in principle be answered. If that computer turns out to be a behaviorist, then it might be a very intelligent computer but it has no consciousness.[8] If, on the other hand, the computer inclines to dualism, then that is proof that it is—that is, he is—perhaps not so intelligent but surely conscious.

*The Consequences for Behavior and Evolution*

The role of consciousness in life can now be generalized. Regardless of the content of Popper and Eccles' (1977) book about the mystery of consciousness, that they wrote it testifies that their consciousness affected their behavior. Likewise, authors like Levin (1986), Tye (1986), and Churchland (1985), whilst denying that there is anything more to behavior than physics, nevertheless referred time and again to the fact that people claim that there is an additional conscious element in their own minds. The conclusion that follows is admittedly a very ambitious one: if this introspective activity indicates that consciousness interferes with behavior, we can extrapolate from these instances to any behavior and say that consciousness is an *additional* cause, although perhaps weak, mingled with the other causes of behavior.

This may seem too daring a conclusion to be derived on the basis of one specific type of behavior, yet we have encouraging support for it from evolutionary theory: earlier we noted that in all physicalistic theories the alleged passivity of consciousness renders the existence of consciousness itself a strange coincidence. On the other hand, the assumption concerning consciousness as an additional factor among the causes of behavior is consistent with Darwinism. We are now in the position to reject the awkward claim (Chase, 1979; Rensch, 1970) according to which evolution and human history would proceed the same way even if all living creatures lacked consciousness. On the contrary, it is more reasonable now to agree with Griffin (1981) that consciousness played a role in evolution. It evolved in higher organisms

[7]Probably the very phrasing of the problem as the "mind-body problem" helps to divert attention from the mysterious nature of consciousness itself to its relationship with the physical.

[8]Prof. J. Agassi, commenting on this point, asked me whether behaviorists lack consciousness. Well, as the test is put, acknowledging that there is a mind-body problem requires not only consciousness—but intelligence as well.

because it made a real difference in the struggle for survival, producing, for example, more powerful experiences and stronger motivations.

Another welcome consequence of our criterion is that, if it is a sufficient one, we can restore some common-sense beliefs that have so far been banned by most philosophers. We can agree with Jackson (1986), for example, that a color-blind scientist will never know what "red" really is, however informed he or she is concerning optics and vision physiology. We have thus dispensed also with the problem of other minds: while hitherto it was commonly held that one cannot know for sure whether one's fellow person is a conscious being or a mere automaton (Fodor, 1981; Schrödinger, 1964), the critical philosopher (to the extent that he or she really needed it) can now rest assured: there are other minds besides his or her own.

*Mind As Maxwell's Demon*

If behavior is influenced by something not yet comprehended by physics, then a sufficiently detailed analysis of behavior must inevitably show that the interference of this extra-factor occurs on the expense of some physical principles. And the question that immediately follows is: Which of the existing basic notions of physics is to be revised? Naturally, we would wish it not to be a too basic one.

Does the influence of consciousness on behavior challenge the first law of thermodynamics, namely, the law of energy conservation? Popper (Popper and Eccles, 1977) acknowledged that such an influence seems to violate this law, and suggested that the latter may be valid only statistically. However, as Larmer (1986) showed, "unless we suppose that only a relatively small amount of interactions take place between minds and bodies, it is very improbable that energy will be even approximately conserved" (p. 279). Larmer has a point: if consciousness must hide in the narrow statistical domain of tiny "holes" in the conservation law, it is negligible. Moreover, Larmer points out that even the explicit dualist Wigner opposes the statistical interpretation of the first law of thermodynamics.

Averill and Keating (1981) also addressed the question: "Does interactionism violate a law of classical physics?"[9] The laws referred to in this paper are energy conservation and the conservation of linear momentum. The authors claim that these laws are valid even if consciousness interferes with the brain's work. Larmer (1986, p. 281) tried to circumvent the problem by distinguishing two forms of the first law of thermodynamics: i) a strong one, stating that "energy can neither be created nor destroyed", and ii) a weaker one, according to which

[9]"Interactionism" is the dualistic theory that maintains that consciousness, as a non-physical phenomenon, actively interferes with the physical dynamics of the brain.

"in a causally isolated system the total amount of energy remains constant". He then suggested that the interactionist can avoid conflict with observational data if he or she adheres only to the weaker form.

But why, then, did these authors not notice that, if energy conservation is the *first* law of thermodynamics, there is also a second one? I intend to show that the confrontation of the active role of consciousness with this law, according to which entropy (disorder) must increase in closed systems, will prove to be a sharper thorn in the side of physicalism.

Sperry (1969, 1980; see also Stapp, 1982) suggested that consciousness manages to affect the brain without interferring with its activity. Consciousness, according to Sperry, does not intervene in the events in the brain, but rather supervenes them. It merely directs neural discharges, their energy being supplied by the organism itself.

Probably without realizing it, Sperry invoked an old ghost known to physicists as Maxwell's demon. In a famous thought-experiment, Maxwell suggested that a tiny being may be able to violate the second law of thermodynamics by opening and closing a hole in a partition dividing a vessel of gas (Bennett, 1987; Ehrenberg, 1967). This way, by accumulating fast molecules in one half of the vessel, thus increasing order and concentrating heat without dispersing significant amounts of energy, the second law may be violated. Sperry's consciousness is supposed to perform the same trick as Maxwell's demon. However, since this demon has long been exorcised, its trick too became outlawed by physics. Brillouin (in Bennett, 1987; Ehrenberg, 1967; Tribus and McIrvine, 1971) has shown that the demon would need information in order to supervene the motion of the molecules. The acquisition of this information requires energy dispersion (e.g., light). Moreover, in order to make sufficient observations the demon must purge the results of earlier observations from his memory—again dispersing energy and increasing the overall entropy. Energy and information thus became closely interrelated notions, as it became clear the both the acquisition of information and its transmission require appropriate exchanges of energy in parallel.

Suppose now that consciousness interferes with the brain processes in the subtle way as Maxwell's demon was supposed to do. In other words, it does not introduce additional energy but directs the brain processes using their already-existing energy. In order to do that, consciousness must focus its influence on *very precise locations* in the brain (such as specific synapses), at *very precise moments*. This enormous precision requires, we now realize, a similarly high number of energy-exchanges between consciousness and the brain.

But can Sperry's model be saved by the recent works concerning the energy-limits of computation? Bennett (1987), Bennett and Landauer (1985), and Landauer (1986) argued that computation can theoretically be carried out with zero energy. These analyses, however, apply only to idealized, reversible com-

puters, requiring a zero temperature. The brain is obviously different from such devices. Moreover, in order to make any use of the results of such a computation, the stages that must precede and follow it, namely the insertion of the input and the reception of the output, involve entropy increase. Consequently, the issue of reversible computation fails to avoid the clash with physics following from the rejection of the passivity dogma: if consciousness interferes with brain-processes, then either the first or the second law of thermodynamics must be violated.

*Afterthoughts*

With this I wish to summarize. I believe that what you, dear reader, and I are doing, namely, brooding over the mind-body problem, is an example of an instance in which a pure action of consciousness, though weak, manifests itself in our observable behavior. From this instance a similar influence can be inferred about behavior. Therefore, any theory relying only on the currently known physical principles for the explanation of human behavior is in principle incomplete. This conclusion, though hard to digest within the present framework of physics, fully explains the evolutionary significance of the development of consciousness. Concerning the question as to which of the physical notions may have to be revised if this argument is valid, I opt for the conservation law as the principle that should be saved, while the second law of thermodynamics should be revised.

This gives rise to many new questions avoided so far. If consciousness is a factor making observable physical effects, will the future progress of physics enable a satisfactory theory of consciousness, or shall we always have to repeat with resignation du Bois-Reymond's dictum *ignoramus et ignorabimus* (we do not know and we will not know)? I believe the former possibility to be the case. Elsewhere (Horwitz, Arshansky, and Elitzur, 1988) a new physical theory has been proposed, in which it was claimed that consciousness is closely related to another phenomenon, equally mysterious, equally self-evident yet equally absent from any physical description, and equally often dismissed as illusion. In short, I refer to the passage of time. Perhaps physics' inability to describe consciousness is a consequence of its geometrical, time-invariant formalism, that ignores this most obvious characteristic of time. This is an extension of a broader novel theory (see Horwitz, 1983, and other references therein) in which spacetime itself is subject to evolution, parametrized by a higher time called $\tau$. This new feature of time was shown to clarify some persistent problems related to quantum mechanics as well as to human perception. In a book in preparation (Elitzur, 1989), a theory along these lines is developed in a greater detail. But, regardless of the proposed answers, the creation of new problems is itself important. By refuting the "whitewashing" theories that

seek cheap compromises between materialism and dualism, we enliven the controversy about the most fundamental mystery; and, to the extent that we have established that consciousness is active, we make it a genuine problem for physics.

## References


Averill, E., and Keating, B.F. (1981). Does interactionism violate a law of classical physics? *Mind, 90,* 102–107.

Bennett, C.H. (1987). Demons, engines, and the second law. *Scientific American, 257*(5), 88–96.

Bennett, C.H., and Landauer, R. (1985). The fundamental physical limits of computation. *Scientific American, 253*(1), 38–46.

Bohm, D. (1951). *Quantum theory.* Englewood Cliffs, New Jersey: Prentice-Hall.

Chase, R. (1979). The mentalist hypothesis and invertebrate neurobiology. *Perspectives in Biology and Medicine, 23,* 103–117.

Churchland, P.M. (1985). Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy, 132,* 8–28.

Cuda, T. (1985). Against neural chauvinism. *Philosophical Studies, 48,* 111–127.

Culbertson, J.T. (1977). The spacetime structure of mental images, Part II. *Psychoenergetic Systems, 2,* 225–254.

Darden, L. (1983). Review of Griffin's *The Question of Animal Awareness. British Journal for Philosophy of Science, 34,* 399–403.

de Beauregard, O.C. (1980). CPT invariance and the interpretation of quantum mechanics. *Foundations of Physics, 10,* 513–530.

Dewan, E.M. (1976). Consciousness as an emergent causal agent in the context of control system theory. In G.G. Globus, G. Maxwell, and I. Savodnik (Eds.), *Consciousness and the brain* (pp. 179–198). New York: Plenum.

Dixon, N.F. (1981). *Preconscious processing.* New York: Wiley & Sons.

Dobzhansky, T. (1967). *The biology of ultimate concern.* New York: New American Library.

Eccles, J.C. (1953). *The neurophysiological basis of mind.* Oxford: Clarendon.

Eccles, J.C. (1986). Do mental events cause neural events analogously to the probability field of quantum mechanics? *Proceedings of the Royal Society, London, B 227,* 411–428.

Eddington, A.S. (1928). *The nature of the physical world.* Cambridge: Cambridge University Press.

Ehrenberg, W. (1967). Maxwell's demon. *Scientific American, 217*(11), 103–110.

Elitzur, A.C. (1989). *Persistent wonders: The ageold riddles of transient time and consciousness.* Unpublished manuscript, Weizmann Institute of Science.

Fodor, J.A. (1981). The mind-body problem. *Scientific American, 225*(1), 114–123.

Folse, H.J. (1985). *The philosophy of Niels Bohr.* Amsterdam: North-Holland Physics Publishing.

Globus, G.G. (1976). Mind, structure, and contradiction. In G.G. Globus, G. Maxwell, and I. Savodnik (Eds.), *Consciousness and the brain* (pp. 269–293). New York: Plenum.

Griffin, D.R. (1981). *The question of animal awareness: Evolutionary continuity of mental experiences* (rev. ed.). New York: Rockefeller.

Horwitz, L.P. (1983). On relativistic quantum theory. In A. Van der Merwe (Ed.), *Old and new questions in physics, cosmology, philosophy, and theoretical biology. Essays in honor of Wolfgang Yourgrau* (pp. 199–187). New York: Plenum.

Horwitz, L.P., Arshansky, R.I., and Elitzur, A.C. (1988). On the two aspects of time: The distinction and its implications. *Foundations of Physics, 18,* 1159–1193.

Jahn, R.G., and Dunne, B.J. (1986). On the quantum mechanics of consciousness, with applications to anomalous phenomena. *Foundations of Physics, 16,* 721–772.

Jackson, F. (1986). What Mary didn't know. *Journal of Philosophy, 133,* 291–295.

Landauer, R. (1986). Computation and physics: Wheeler's meaning circuit? *Foundations of Physics, 16,* 551–564.

Larmer, R. (1986). Mind-body interaction and the conservation of energy. *International Philosophical Quarterly, 26,* 277–285.

Levin, J. (1986). Could love be like a heatwave? Physicalism and the subjective character of experience. *Philosophical Studies, 49,* 254–261.
McMullen, C. (1985). "Knowing what it's like" and the essential indexical. *Philosophical Studies, 48,* 211–233.
Mott, P.L. (1982). On the function of consciousness. *Mind, 91,* 423–429.
Natsoulas, T. (1988). On the radical behaviorist conception of pain experiences. *Journal of Mind and Behavior, 9,* 29–56.
Penrose, R. (1987). Quantum physics and conscious thought. In B. Hiley and D. Peat (Eds.), *Quantum implications: Essays in honour of David Bohm* (pp. 105–120). New York: Methuen.
Peres, A. (1986a). When is a quantum measurement? *American Journal of Physics, 54,* 688–692.
Peres, A. (1986b). Existence of "free will" as a problem of physics. *Foundations of Physics, 16,* 573–584.
Popper, K.R. (1982). *Quantum, theory and the schism in physics.* (Part III of the Postscript to *The logic of scientific discovery.* W.W. Bartley, Ed.). London: Hutchinson.
Popper, K.R., and Eccles, J.C. (1977). *The self and its brain.* Berlin: Springer Verlag International.
Putnam, H. (1964). Minds and machines. In A.R. Anderson (Ed.), *Minds and machines* (pp. 72–97). Englewood Cliffs, New Jersey: Prentice-Hall. (Originally published 1960)
Putnam, H. (1975). *Philosophical papers Vol. 2: Mind, language and reality.* Cambridge: Cambridge University Press.
Rensch, B. (1970). Evolution of matter and consciousness and its relation to panpsychistic identism. In M.K. Hecht and W.C. Steere (Eds.), *Essays in evolution and genetics in honor of Theodosius Dobzhansky* (pp. 97–119). Amsterdam: North-Holland Publishing Company.
Rucker, R.v.B. (1982). *Infinity and the mind: The science and philosophy of the infinite.* Boston: Birkhouser.
Schmidt, H. (1982). Collapse of the state vector and psychokinetic effect. *Foundations of Physics, 12,* 565–581.
Schrödinger, E. (1958). *Mind and matter.* Cambridge: Cambridge University Press.
Schrödinger, E. (1964). *My view of the world.* Cambridge: Cambridge University Press.
Shoemaker, S. (1975). Functionalism and qualia. *Philosophical Studies, 27,* 291–315.
Skillen, A. (1984). Mind and matter: A problem which refuses dissolution. *Mind, 93,* 514–526.
Sperry, R.W. (1969). A modified concept of consciousness. *Psychological Review, 76,* 532–536.
Sperry, R.W. (1980). Mind-brain interaction: Mentalism, yes, dualism, no. *Neuroscience, 5,* 195–206.
Stapp, H.P. (1982). Mind, matter, and quantum mechanics. *Foundations of Physics, 12,* 363–399.
Thagard, P. (1986). Parallel computation and the mind-body problem. *Cognitive Science, 10,* 301–318.
Tribus, M., and McIrvine, E.C. (1971). Energy and information. *Scientific American, 224*(9), 179–188.
Turing, A.M., (1964). Computing machines and intelligence. In A.R. Anderson (Ed.), *Minds and machines* (pp. 4–30). Englewood Cliffs, New Jersey: Prentice-Hall. (Originally published 1950)
Tye, M. (1986). The subjective qualities of experience. *Mind, 377,* 1–17.
Wassermann, G.D. (1983). Quantum mechanics and consciousness. *Nature and System, 5,* 3–16.
*Webster's Third New International Dictionary* (1981). Springfield, Massachusetts: Merriam-Webster.
Wheeler, J.A. (1980, June 5). *Delayed-choice experiments and the Bohr-Einstein dialog.* Paper presented at the joint meeting of the American Philosophical Society and the Royal Society, London.
Wheeler, J.A. (1981). Not consciousness but the distinction between the probe and the probed as central to the elemental quantum act of observation. In R.G. Jahn (Ed.), *The role of consciousness in the physical world* (pp. 87–111). AAAS Selected Symposium 57. Colorado: Westview Press, Boulder.
White, N.P. (1985). Prof. Shoemaker and the so-called "qualia" of experience. *Philosophical Studies, 47,* 369–383.
Wigner, E.P. (1970). Physics and the explanation of life. *Foundations of Physics, 1,* 35–45.
Wigner, E.P. (1975). *Symmetries and reflections: Scientific essays.* Bloomington: Indiana University Press.
Wilkes, K.V. (1984). Is consciousness important? *British Journal for the Philosophy of Science, 35,* 223–243.
Wilson, D.L. (1976). On the nature of consciousness and of physical reality. *Perspectives in Biology and Medicine, 19,* 568–581.