COMPUTATION AND CONSCIOUSNESS*

> Their favorite target was the rigid and silent Olympia, who, her beautiful appearance notwithstanding, was assumed to be hopelessly stupid, which was thought to be the reason Spalanzani had kept her so long concealed. Nathanael heard all this, not without inner fury, but he said nothing. "What would be the use," he thought, "of proving to these fellows that it was their own stupidity which precluded them from appreciating Olympia's profound and beautiful mind."
>
> from E. T. A. Hoffmann's *The Sandman*

THERE is no more amazing and puzzling fact than that of consciousness. Little wonder, then, that Philosophia, daughter of Thaumas, should wile away so many hours speculating on the nature of the mind. For, as Thomas Nagel has noted, it is consciousness that makes the mind-body problem intractable, it is the unfathomable gap between physical process and subjective awareness which mocks our search for the filaments that bind the corporeal and the mental together. Any program that holds promise for solving the mind-body problem deserves our closest attention. I wish to focus on one such program in this essay.

There is little doubt that humans have a mental life, because we have brains. Yet granted that brain activity somehow supports mental activity, the question still arises: In virtue of what do brains support minds? Which properties of our gray matter are essential, and which accidental, for our mentality? What level of description or abstraction distills the essence of mental life? The answer to this question is the beginning of wisdom in the philosophy of mind.[1]

* An embryonic version of this paper was presented at the University of Pittsburgh in January 1988. I am particularly indebted to Kevin Kelly, Clark Glymour, and Peter Spirtes for forcing me to clarify the main points at issue. I do not know that any of them would endorse any part of the argument. I am sure that they would all still find Olympia to be unpleasantly baroque, but I am afraid that, like poor Nathanael, I have fallen in love with her.

[1] One might wonder why the answer is not also the end of wisdom in the philosophy of mind. In this paper, I mean only to consider that part of a solution of the mind-body problem which purports to detail necessary and sufficient conditions for physical systems to have minds. It might be possible to discover such conditions and still be puzzled about how subjective states arise from certain physical conditions. Making the connection between mind and body intelligible, seeing how subjective states can be brought about by objective conditions, is a further problem, and of another order of magnitude.

0022-362X/89/8608/407–432

One clear answer to this question is now being widely championed: the *computational* structure of the brain is what bestows mental properties. We must abstract away the particular physical, biochemical, and neural features of brains to see what really makes them tick. The computational or information-theoretic description captures what is of importance.

The computational approach has been most intensively applied to problems of intentionality, language use, and representation. But, if it is to untie the Gordian knot uniting *res cogitans* with *res extensa*, the computational paradigm must eventually directly face the problem of consciousness. So we must ask: Is a computational theory of consciousness possible? Or, given the centrality of consciousness to the mind-body problem, we may equally phrase the question: Is a computational theory of mind possible?

Before considering the implications of a computational theory of consciousness, let me pause to define our terms more clearly. By 'consciousness' I mean *subjective phenomenal states* or *modes of awareness*. The most obvious examples of conscious episodes are sensory: tickles, pains, visual experiences, and so on. But they are not confined to straightforwardly sensory events. There is a certain phenomenology associated, for example, with my pondering the fact that ice is made of water. In such cases, the phenomenal properties of the experience need not determine the content of the proposition entertained. If a causal theory of reference is correct, a molecule-for-molecule identical replica of my brain, if just brought into existence, may not be capable of entertaining the proposition that ice is made of water. Still, our best guess is that such a brain would support identical states of consciousness to mine, identical phenomenal states. In Nagel's terms, what it would be like to be the person with that brain would be just what it is like to be me. The content and structure of those subjective states of awareness are what I mean by 'consciousness'.

Of course, the thesis that physically identical brains would support phenomenally identical states of consciousness is not analytic. But some such physicalist assumption underlies all contemporary research into perception and neuro-physiology. Furthermore, it seems to be an essential thesis for the computationalist. For computational structure supervenes on physical structure, so physically identical brains are also computationally identical. Hence, any mental property that can be given a purely computational analysis ought to be shared by physically identical brains.

In the sequel, a somewhat stronger claim about supervenience shall be employed. States of awareness and sensory events take place

in time; they are fairly precisely datable. One can assert that Sam had a toothache at 12:05 or that Sheila spent five minutes wondering about Fermat's last theorem. A natural, indeed nearly inescapable, explanation for this is that conscious events and episodes supervene on concurrent physical events and processes. One's phenomenal state at a time is determined entirely by one's brain activity at that time. Hence, two physical systems engaged in precisely the same physical activity through a time will support the same modes of consciousness (if any) through that time. Let us call this the *supervenience thesis*.

Again, this is a substantive thesis, albeit one for which we have some fairly direct evidence. We know that we can induce pains and visual experiences at a time by stimulating particular areas of the brain at that time. We know that particular types of mental activity are directly correlated with concurrent types a brain activity as revealed, for example, by the electro-encephalograph. It is not a far leap to suppose that all modes of subjective awareness supervene on that brain activity.

It may be useful to intimate exactly how the supervenience thesis will be used in the sequel. Suppose that a system exists whose activity through a period of time supports a mode of consciousness, e.g., a tickle or a visual sensum. The supervenience thesis tells us that, if we introduce into the vicinity of the system an entirely inert object that has absolutely no causal or physical interaction with the system, then the same activity will still support the same mode of consciousness. Or again, if the activity of a system supports no consciousness, the introduction of such an inert and causally unconnected object will not bring any phenomenal state about.

The plausibility of this particular application of the supervenience thesis derives from the unity and completeness of occurent phenomenal states. A pain has its entire existence and being in the event of its being perceived, and its perception is a single, unified occurrence. This suggests that the supervenience space of the pain must also have a sort of unity or connectedness, presumably a causal connectedness. If an active physical system supports a phenomenal state, how could the presence or absence of a causally disconnected object effect that state? How could the object enhance or impede or alter or destroy the phenomenal state except via some causal interaction with the system? Since the phenomenal state is entirely realized at the time of the experience, only the activity of the system at that time should be relevant to its production. The presence or absence of causally isolated objects could not be relevant. This is all the supervenience thesis needs to say.

Having stated both what I mean by 'consciousness' and the supervenience relation that I take to hold for it, let us now turn to 'computational'. What is a computational theory of consciousness?

To avoid the quagmire of behaviorism, one must posit that the (actual and counterfactual) input-output relations of the brain do not alone determine its mental properties. The internal structure of the brain is essential: one must examine how it generates the output from the input. Broadly speaking, any program that is concerned with characterizing the internal structure of this brain activity may be denominated a form of *functionalism*. Functionalists examine how the brain is organized, how it does what it does.

If functionalism is so broadly defined, though, it becomes nearly vacuous. One of the "ways" that the brain does what it does is by being a neuro-physiological object, by transmitting nerve impulses, etc. One of the ways the brain is organized is its chemical organization, and its structure can be defined at the level of physical structure. If the level of organization or degree of abstractness of a functional theory of mind is not constrained, then it is bound to succeed, at least in giving sufficient conditions for consciousness. This will be guaranteed if physicalism is true. To give functionalism any real bite, then, some precise specification of the appropriate level of functional organization must be set. One exact solution to this problem is proposed by the computational or information-processing theories of mind. They posit that the important level of functional organization of the brain is its computational structure.

The fundamental notion of a computational description is that of a machine table. The table describes how the internal states of the machine are subjunctively connected to one another and to various input and output mechanisms. In the case of a Turing machine, the example we shall be concerned with, the machine table specifies how the machine in a given state would respond to an entry on the machine's tape by (a) changing the entry on the tape, (b) moving to another address on the tape, and (c) going into a new internal state. The machine table thus determines exactly how the state of the machine and tape will evolve given any input and initial state.

The foremost advantage of a computational description of a system is its precision. Once the internal states and proper operating parameters of a physical system have been specified, one can determine by straightforward physical analysis how the states would be subjunctively connected and hence what the appropriate machine table is. One can compare the machine tables of any two systems to see if they are the same, and thereby determine whether they have the same program or are, at a given time, performing the same

computation. The computational approach gives a determinate content to functionalism by precisely specifying the level of abstraction at which a physical system should be analyzed.

We should remark just how high a level of abstraction a computational description employs. The machine table does not directly specify what sort of physical or material properties a system running a program must have. Nor can much in the way of physical constraints be even indirectly derived from the machine table. We may contrast, by way of example, the functional description of a valve lifter. Jerry Fodor[2] has used the concept of a valve lifter as an archetypical functional concept. From the fact the something is a valve lifter (as opposed to, say, a camshaft), one cannot infer its precise material structure or composition. To be a valve lifter is to play a certain functional role in the internal economy of an engine. Still, there seem to be some physical constraints on valve lifters: they must, for example, transmit enough energy to lift a valve. A weak photon cannot be a valve lifter or (to use another functional type) a mousetrap, because it cannot provide the physical wherewithal to lift a valve or catch a mouse. These physical constraints derive from the fact that the functions involved are ultimately specified by reference to physical objects and properties. Engines must produce mechanical work, mousetraps must immobilize a given type of animal. The items playing certain functional roles in service of these ends may have to meet some physical standards in order to achieve them. So, although these functional concepts can abstract from the exact specification of material structures, still some fairly strong physical capacities may be implied by them.

In contrast, the terms of a computational description involve no ultimate physical or mechanical goal to be achieved. The only physical requirements that a system must meet in order to instantiate a certain machine table are that (1) there must be at least as many physically distinguishable states of the system as there are machine states in the table, (2) the system must be capable of reacting to and changing the state of the tape, and (3) there must be enough physical structure to support the subjunctive connections specified in the table. Similar remarks apply to the tape itself.[3]

[2] *Psychological Explanation* (New York: Random House, 1968), pp. 113 ff.

[3] If brains support mental states in virtue of their computational structure, then it is clear that minds could be realized in almost any stable material substrate. The intuition that persons may only accidentally be made of flesh and blood dates back at least to Socrates the Younger (cf. Aristotle, *Metaphysics* 1036b20 ff.), and is part of the credo of the AI community. Witness the strong reaction to John Searle's suggestion that "whatever else intentionality is, it is a biological phenomenon, and it

A computational theory of consciousness must therefore hold that particular physical systems are capable of consciousness because they can be correctly described as machines that have been appropriately programmed. That is, there must be some identification of internal states and parameters of normal operation such that, under that interpretation, the machine instantiates an appropriate machine table, and any system that satisfies that table will be capable of consciousness. Of course, there will presumably be an infinite number of such programs that would bestow the capacity for consciousness.

Furthermore, if we move from the capacity for consciousness to its exercise, a computational theory of mind must hold that a system actually is conscious, is supporting an occurrent mode of awareness, when it is actually running the appropriate program on the appropriate tape. So, if for the last five minutes I have had a toothache, or the subjective experience of pondering Fermat's last theorem, then there must be some program and some state of the machine tape such that anything that runs that program on that tape (starting in a particular machine state and progressing through a finite number of cycles) will create or support or underlie or be associated with an occurrence of that mode of consciousness. Finally, a computational theory of consciousness must hold that a necessary condition for supporting consciousness be that a system be describable as a nontrivial computational system performing a nontrivial computation. Otherwise, a system with no interesting computational structure could be conscious, so computation could not be essential to consciousness.

If we cast these last two requirements in terms of Turing machines,[4] we have the following two principles. Take any type of phenomenal conscious state ϕ:

> *Sufficiency Condition:* There must be some program π and some state of the machine tape τ and some sequence of machine states $S_{[0]}$, $S_{[1]}$, $\cdots S_{[N]}$ such that any machine programmed with π, operating on τ, and performing the computation $S_{[0]}$, $S_{[1]}$, $\cdots S_{[N]}$ (in accordance with π) will, during the time that the computation is taking place, support the mode of consciousness ϕ.
>
> *Necessity Condition:* A necessary condition for a physical system to support ϕ is that it be describable as a Turing machine performing some nontrivial computation.

is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomenon" ("Minds, Brains and Programs," *The Behavioral and Brain Sciences III*, 3 (1980): 424). Zenon Pylyshyn, for example, asserts that "we cannot take as sacred anyone's intuitions about such things as whether another creature has intentionality—especially when such intuitions rest (as Searle's do, by his own admission) on knowing what the creature (or machine) is *made of*. . . . Clearly, intuitions based on nothing but such anthropocentric chauvinism cannot form the foundation of a science of cognition" (*ibid.*, p. 443). As we will see, there are other levels of description less abstract than the computational level which also sustain the possibility that the functional organization can be realized in almost any material.

[4] Similar conditions can easily be stated for finite-state machines. I take the Turing machine as the most general case.

(The subscripted number in square brackets denotes the i^{th} step in this particular computation. So this computation starts in state $S_{[0]}$ and with the read/write head at location $T_{[0]}$, next moves to $S_{[1]}$ and $T_{[1]}$, and so on. Subscripts without brackets indicate a general numbering of the different machine states and tape addresses. So, if in the course of the computation, the second tape address is visited at the 5^{th} and 18^{th} steps, $T_{[5]} = T_{[18]} = T_{[2]}$.)

The burden of the remainder of this essay will be to demonstrate that the sufficiency condition, the necessity condition, and the supervenience thesis form an inconsistent triad, and hence that an acceptable computational theory of consciousness is not possible. Let me first, however, try to forestall some possible misunderstandings.

I. WHAT THIS ARGUMENT IS NOT

In developing the argument that follows, I shall construct a machine that will run the program π. The machine will be somewhat peculiar and complicated, but its purpose should not be misconstrued. The point of the machine is to demonstrate the inconsistency of the three conditions listed above. This line of argument must be distinguished from another common dialectical gambit which may be called the *ploy of funny instantiation.*

The ploy of funny instantiation, in its crudest form, operates by first pointing out that any Turing machine can be realized in some peculiar materials (e.g., water pipes or toilet paper or windmills and old beer cans), and then appealing to intuitions that that sort of stuff just cannot be conscious or have mental states or intensionality. I fully reject this ploy, because I approach the mind-body problem as one thoroughly mystified by the relationship between physical activity and consciousness. Although I may have intuitions that windmills and old beer cans, no matter how cleverly connected, cannot be conscious, still I cannot think of one reason to accord those intuitions any weight. How pulses of water in pipes might give rise to toothaches is indeed entirely incomprehensible, but no less so than how electro-chemical impulses along neurons can. For all I know, a conscious state may equally well be created by either of these sorts of process.

John Searle's Chinese room argument is a sort of funny instantiation ploy, but with a twist.[5] Searle does not directly argue that this particular system, a human shuffling some cards about, is not the kind of thing which can be conscious or have intensionality or understanding. Rather, since the man is just the kind of thing which is conscious, Searle simply presupposes that whatever intensionality or understanding arises by shuffling the cards would have to belong to the same mental entity, the same self, that is associated with the brain activity of the man in the room. Since *that* personality does not come to understand Chinese, Searle assumes that nobody or nothing understands Chinese. But the inference does not go through.

The problem is perhaps clearer in the case of sensations. Suppose that someone claims that the operation of a particular program will always create or support the experience of a toothache. If we allow the man in the toothache room to be our Turing machine and perform the appropriate computational activity, we do not suppose that thereby his tooth would start to hurt. Rather, a phenomenal state would supervene on his activities, a phenomenal state disjoint from his consciousness. For, on this theory, his consciousness is created by the pattern of activity of his neurons, not by the computational structure of the manipulation of the cards. There is no reason to insist that the phenomenal states associated with the former must combine with the phenomenal states associated with the latter into one self.

This is exactly the solution to Searle's problem offered by several of the commentators on his paper. Robert Wilensky writes: "Searle's mistake of identifying the experiences of one system with those of its implementing system is one philosophers often make when referring to AI systems" (*ibid.*, p. 450). Aaron Sloman and Monica Croucher suggest that an appropriate analogy to the relationship between the consciousness of the man in the room and the consciousness created by the man in the room may be found in cases of multiple personalities (*ibid.*, p. 448). And Marvin Minsky directly asserts that "in the case of a mind so split into two parts that one merely executes some causal housekeeping for the other, I should suppose that each part —the Chinese rule computer and its host—would then have its own separate phenomenologies—perhaps along different time scales" (*ibid.*, p. 440). (Note that, although Searle's argument is concerned with intensionality and understanding, these writers also speak of consciousness and phenomenologies, implying that they accept a computational theory of consciousness.) Searle has done nothing

[5] The Chinese room was constructed in Searle, *op. cit.*

to discount the possibility of simultaneously existing disjoint mentalities.

The time-scale problem underlies another sort of funny instantiation ploy. The clock of a computer can be slowed down indefinitely without affecting its computational structure. A machine might, for example, go through only one cycle every 1,000 years. It might take billions of years to complete the computations from $S_{[0]}$ to $S_{[N]}$. It might spend 999 years out of each 1,000 disassembled, parts widely scattered, only being put together for the single millennial cycle. None of these facts would affect its computational description, so, if we accept the sufficiency condition, there would still be associated with the billion year process a conscious episode of type ϕ. This does seem exceedingly odd, but I do not see any way of promoting this oddness into a rational objection to the sufficiency condition.

I am not, then, constructing this machine in order to appeal to primitive intuitions about what sort of materials can or cannot support consciousness. The question I want to ask is rather the following: Given that consciousness supervenes on physical processes and activity, what is the minimum level of physical activity that a machine must perform in order to run program π from $S_{[0]}$ to $S_{[N]}$, and hence (according to the computational theory) to support phenomenal state ϕ? What is the laziest machine that can be conscious? This problem shall guide our construction.

II. THE RULES OF THE GAME

In order to run π from $S_{[0]}$ to $S_{[N]}$, the following three conditions must be fulfilled. First, the machine itself must pass sequentially through N machine states. Second, the read/write head of the machine must sequentially visit tape locations $T_{[0]}$ to $T_{[N]}$. In doing so, it may have to change the data entries of any or all locations. These first two conditions contain all of the activity that can be explicitly derived from the definitional conditions of what it is to run π on τ. But there are other conditions that may somehow implicitly demand that much more complicated activity is needed. For, thirdly, the physical state of the machine must also support all of the counterfactuals inherent in the machine table. Not only must the machine actually go from $S_{[0]}$ to $S_{[1]}$ given that the data location $T_{[0]}$ actually contains, say, a 0, but it must also be constructed so that it would have gone into $S'_{[2]}$ (in accordance with π) had $T_{[0]}$ contained a 1. And the counterfactuals ramify geometrically: for every state it might have gone into, the program will specify where it would have gone after that (for each possible input), and so on. Thus, even though the machine may go into only a small number of states in running π from $S_{[0]}$ to $S_{[N]}$, still in order to be running π it may have to have a tremendous number of

states accessible to it, all interconnected in exactly the right way. This at least prima facie suggests that the physical activity needed to run π must be immensely more complicated, and the transitions from one state to another immensely more complex, than the two explicit conditions on physical activity suggest. Our machine will show that this plausible argument is not correct: the physical activity involved in running a program may be limited to the most trivial means of fulfilling the explicit conditions, with the satisfaction of the third condition on counterfactual structure requiring almost no additional activity at all.

In constructing our lazy machine, certain tricks shall be ruled out of order. For example, in imagining how little physical activity is needed to pass from some state S_i to another one S_j, someone might suggest that no activity is needed. Let a rock sitting on a table be the machine. Now let S_i be: sitting on the table from 12:00 to 12:01. Let S_j be: sitting on the table from 12:01 to 12:02. The machine will effect a transition between the two states without undergoing any physical change at all. I shall take such tricks to be inadmissible; it is in the spirit of computationalism that the machine states be physically meaningful states and that the state transitions be physical activities. I shall also require that the tape addresses be physically distinct entities and that the informational state of an address (0 or 1) be specifiable in terms of its intrinsic physical state. Lastly, the state of the machine must be defined without reference to the state of the tape, and vice versa. Even with these restrictions, we shall see that any program can be run with an uncomfortably small amount of physical activity. Without them, the activity could be reduced even further.

III. THE CONSTRUCTION

In constructing our machine, I first posit that we already have a machine programmed to run π. Let us call this machine *Klara*. Klara can in principle be constructed in whatever manner and out of whatever materials one prefers. She may be silicon or protein, electrochemical or hydraulic. I shall imagine her as a clockwork mechanism, but nothing turns on this choice. Klara may well be unimaginably complex; as a clockwork she may have to be as large as the solar system. These details will ultimately be irrelevant to the construction.

Although Klara's own composition will be subject to no constraints, Klara's tape is to have a very particular form. She will store and retrieve information in a (potentially) infinite series of water troughs. Each trough can store one binary bit of information by being either full of water or empty. We can imagine Klara's read/write head getting information from an address by dropping a float

down into the trough to determine whether it is full or not. Entries can be changed by either filling an empty trough or opening a drain on a full one. To begin with, then, a section of Klara's tape would be as depicted in figure 1.

Now, certain sorts of manipulation can be performed on Klara's tape which (with compensating adjustments) would in no way affect her computational structure. We may, firstly, arbitrarily alter the order of any finite number of the troughs, so long as we provide Klara with a map of the alterations. Before this change, if Klara is at location 18 and has been instructed to go to the location +2, she will just move two troughs to the right. After the rearrangement, she would first determine that she is at address 18, calculate that she must go to address 20, locate that trough on the map, and move her read/write head there. So the spatial sequence of the troughs need not reflect their "computational sequence." We may so contrive that any sequence of addresses lie next to each other spatially.

Secondly, any tape address can be *bi-located.* If it is convenient for some reason to have a single address be accessible from two different spatial locations, we need only set up a trough at each location and connect them with a pipe. The two now jointly form one address, with its informational content available from each location. The address can also have its contents changed, i.e., be filled or emptied, by means of an operation performed at either trough. Again, information about such bi-location (or more highly multiple location) can be encoded in Klara's map.

Aside from these possible manipulations of the tape, we need introduce only one more item. A *block* is a device that can halt the operation of Klara and freeze her in a particular machine state. If she is clockwork, we can imagine placing a simple block of wood between

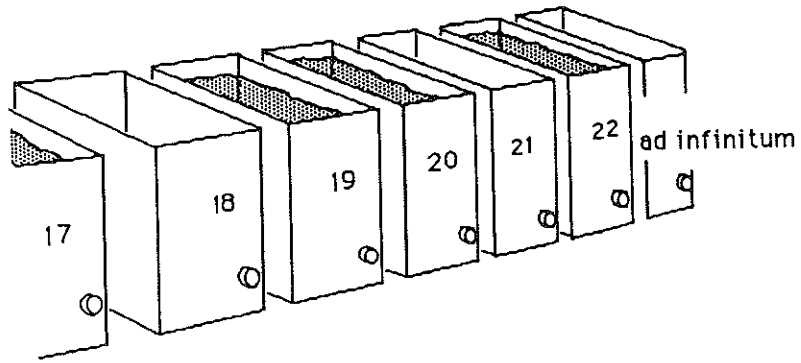

Figure 1: Klara's Original Tape

her main operating gears, immobilizing them. The gear teeth of a blocked machine may well not even physically contact each other, remaining suspended motionless in space. Unblocking a machine allows it to continue running as normal.

The first step in our construction is to rearrange Klara's tape so that addresses $T_{[0]}$ to $T_{[N]}$ lie spatially in sequence, $T_{[0]}$ next to $T_{[1]}$ next to $T_{[2]}$, etc. (Recall that the numbering in square brackets, $T_{[n]}$, indicates the trough visited at step n of the computation.) Given the possibility of arbitrarily reordering and bi-locating troughs, this must be possible. We start with $T_{[0]}$, move $T_{[1]}$ next to it, and so on. If $T_{[i]} = T_{[j]}$ for some $i < j$, so the program requires the machine to "go back" to an address already visited, we simply bi-locate the address by connecting two troughs with a pipe. Any address can be multiply located to any finite degree. After the rearrangement, the relevant section of tape would look as depicted in figure 2.

Now, when Klara operates π on τ from $S_{[0]}$ to $S_{[N]}$ her read/write head will just sequentially move to the right. This may be accomplished by a very complex operation, by scanning address numbers and performing calculations and referring to maps, but the net result will be that Klara will always move to the next trough along, read it, perhaps fill or empty it, move to the next, etc., until reaching $T_{[N]}$.

We will begin constructing our new machine, Olympia, by contriving the simplest means of performing this sequence of operations on the tape. The task is nearly trivial: we need only an armature designed to travel left to right which will fill or empty the appropriate troughs as it passes by. Emptying might be accomplished by hitting a rod attached to the drain plug of the trough. Filling can be accomplished by a hose attached to the armature. If the trough being passed is to be filled, the water is allowed to flow in. If not, a barrier

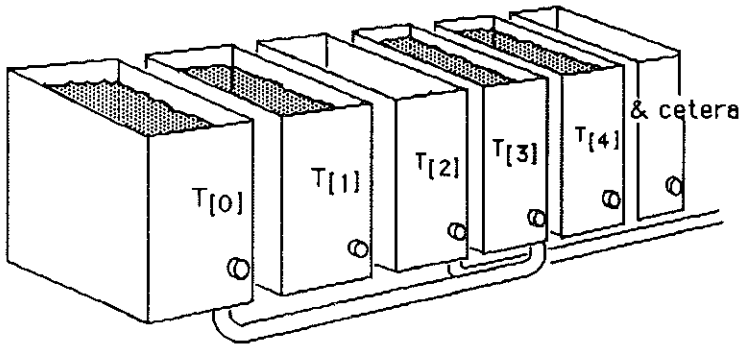

Figure 2: The Reordered Tape

placed at the top of the trough might divert the water away. So far, then, Olympia is just the device of figure 3.

As depicted, Olympia's armature would strike the first rod and empty $T_{[0]}$, leave $T_{[1]}$ unchanged, fill $T_{[2]}$ and $T_{[3]}$ ($=T_{[0]}$, and so at that point empty), etc. So she can already fulfill one of the requirements of a machine operating π on τ: the manipulations of the tape will be done correctly and in the appropriate order. The armature also is doing enough to fulfill a second requirement: the armature passes sequentially through N distinct physical states. One need only define as a distinct state the state of the armature being over a particular trough. If we could identify the state of being over the trough marked '$T_{[i]}$' with the machine state $S_{[i]}$ (for $i = 0$ to N), then Olympia would already be going through the appropriate sequence of machine states as the armature moves, and would already be running π. Of course, one cannot simply by fiat command that the armature position constitutes the appropriate machine state. A particular physical state only becomes interpretable as a machine state of a system programmed with π in virtue of standing in the right counterfactual or subjunctive relations to the tape and to the whole constellation of other states in the machine table for π.

Indeed, as matters stand, it seems impossible to maintain that the armature alone is instantiating any program or is doing any computing at all. Its operations are entirely oblivious to the state of the tape. The sequence of states it passes through and the manipulations it performs are entirely unresponsive to the data stored in the addresses. If one wants to ascribe a computation to it, the computation

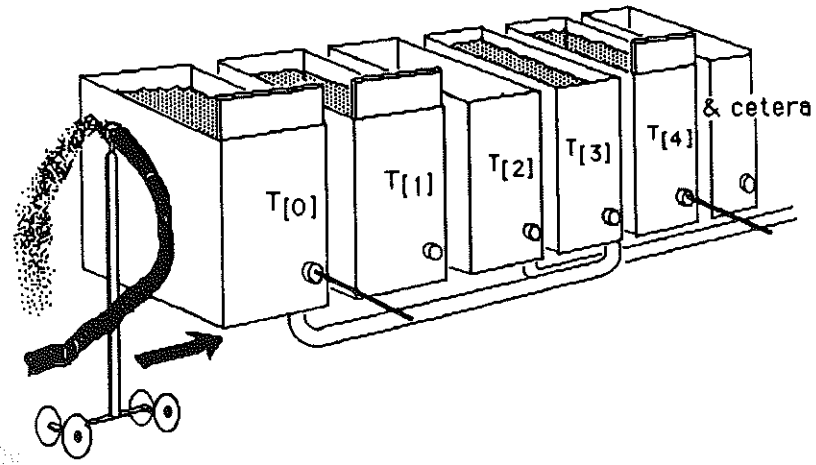

Figure 3: The Armature

would be just a constant function that gives the same output for any input.[6] Such a program would be, in as strong a sense as possible, trivial. It would make a mockery of the computational theory of mind, if the "computation" of a constant function after the manner of the armature could be sufficient to support any mode of consciousness. Indeed, the necessity condition asserts that any conscious entity must be describable as a nontrivial Turing machine running a nontrivial program. If one is to elucidate mental capacities as computational, the computations involved must not be just constant functions, especially constant functions "computed" in such a nonresponsive way. The computationalist must assert, then, that the armature alone, traveling from left to right, does not support a state of consciousness.

The point, again, is not that the armature and tape are the wrong kind of stuff to be conscious. For all I know, appropriately constructed troughs and water hoses could have toothaches, or think about Fermat's last theorem. The point is rather that the armature does not have the right kind of computational or information-processing structure to be conscious according to the computationalist's own thesis. For the armature is not processing any information at all.

As yet, the computationalist has no grounds to be alarmed that the activity of the armature supports no conscious state. The computationalist thesis is committed to maintaining that any system that runs π on τ from $S_{[0]}$ to $S_{[N]}$ is conscious, but the armature clearly does not instantiate π. The vast majority of machine states needed to describe π have not been correlated with any possible physical state of the armature, nor are the counterfactuals implied by π supported by the armature's structure. Still, the armature is performing enough activity to satisfy the explicit demands for physical change entailed by the fact that a machine is running π on τ. All that is lacking is the counterfactual structure needed to support the program. Somehow, the machine needs to have available the entire constellation of machine states that it would have gone into (according to π) had the data entries in the tape been different from those of τ.

The crux of the matter appears, though, when we note that the counterfactuals can be supported by machinery that is physically inert during the computation. That is, although we must add some more structure to Olympia to ensure that she is running π, that

[6] Strictly speaking, to get a constant function for the output in troughs $T_{[0]}$ to $T_{[N]}$, one would have either to place an emptying rod or to remove a barrier from every trough location, depending on whether the trough is respectively empty or full after the associated machine state $S_{[i]}$ has been completed. This addition would have no effect on the construction.

structure need not do anything, need not be physically active, while this bit of π is running. We already have all the physical activity needed to run π.

So long as the troughs are filled according to the specification τ, the armature will perform all of the correct manipulations on it. What we need is some extra structure that would be activated were any of the tape addresses to hold different data. We can achieve this as follows.

At each trough location $T_{[i]}$ we attach a float. The float can be either up or down. Each float is held fixed in place by a latch, and the latch is released only when the armature passes by. Initially, each float is fixed in the position corresponding to the water level in $T_{[i]}$ when the machine goes into state $S_{[i]}$ during a normal run of π on τ. If during such a run address $T_{[i]}$ is full when the i^{th} step is reached, we will fix the float up; if empty, we fix it down. So arranged, during a normal run of π on τ, none of the floats will in fact change position. When the latch is released the float is already set at the appropriate water level and so stays still.[7] Were the program to be run with the tape in a different state τ', in which the content of a trough $T_{[i]}$ differs from that in τ, then the float will either rise or fall when the armature reaches $T_{[i]}$ and releases the latch. So for any tape state that would cause a machine running π to evolve differently than it will running on τ, there will be at least one float that will move when the armature goes by.

The floats have provided Olympia with some subjunctive sensitivity to variations from τ. Now it is just a matter of hooking that sensitivity into the right machinery. To do this, we make $N+2$ copies of Klara. The first copy is blocked in the state $S_{[0]}$, and the read/write head is placed in trough $T_{[0]}$. The second copy is blocked in $S_{[1]}$ and set in $T_{[1]}$, and so on up to $S_{[N+1]}$. The block in the first machine (which is, recall, a small bit of wood jamming the gears) is attached by a chain to the float at $T_{[0]}$, and so on until the machine blocked in state $S_{[N]}$ is attached to the float a trough $T_{[N]}$. Now, if any float should rise or fall, it will unblock the corresponding copy of Klara. The machine blocked in state $S_{[N+1]}$ is set so that it will be set running, if the armature completes its trip from $T_{[0]}$ to $T_{[N]}$. Finally, a return chain ensures that as soon as any of the blocked machines is set running, the armature is immediately shut down.

We now have Olympia in all her glory (see figure 4). We need only give the definition:

[7] The latch is reengaged before the armature itself does anything to affect the level. Only the water level at the moment that the armature arrives at the trough determines the float's behavior.

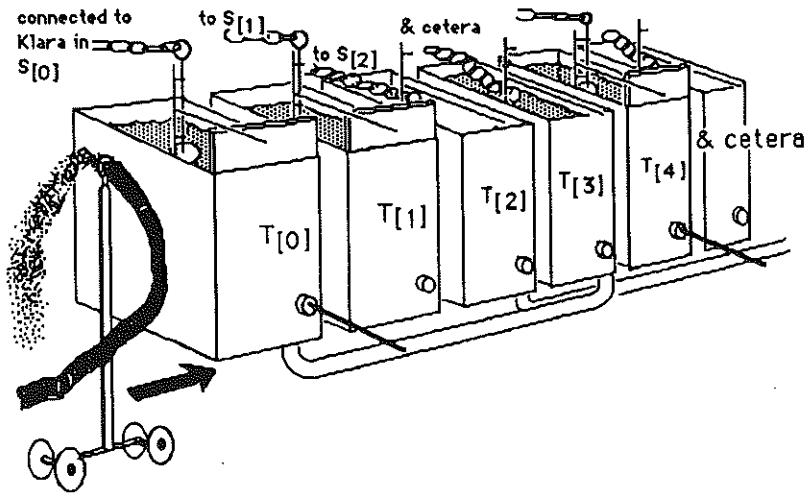

Figure 4: Olympia Unveiled

Olympia is in state S_i iff either one of the copies of Klara is unblocked and is in state S_i or $S_i = S_{[j]}$ for some j and the armature is running and passing over trough site $T_{[j]}$.

A little analysis confirms that Olympia is now programmed to run π. Once one of the copies of Klara gets unblocked, Olympia will run π because Klara is programmed to run π. But even while no copy of Klara is unblocked, while only the armature is moving, Olympia still instantiates π. For as they are now defined, the machine states are tied together by the right subjunctive relations. If we start Olympia running on τ she will, by construction and according to the definitions, pass through exactly the sequence of machine states and perform exactly the tape manipulations demanded by π. And, if we start Olympia running with the tape in a state other than τ, then at exactly the point that the armature alone would go wrong (where the data entry differs from τ), one of the copies of Klara wil be unblocked to take over the task. Thus, no matter what the state of the tape, the sequence of machine states as defined and the sequence of manipulations of the tape will be in exact accordance with π. But being programmed with π is nothing over and above having just these dispositions.

IV. THE BIND

We are now in a position to demonstrate the conflict that arises among our three theses. According to the sufficiency condition, when Olympia runs on τ and operates π from $S_{[0]}$ to $S_{[N]}$, she is

conscious and experiencing phenomenal state ϕ. But, when Olympia runs on τ, all that happens physically is that the armature tracks from left to right, emptying and filling troughs and momentarily unlatching floats. In particular, absolutely no physical activity occurs in any of the blocked copies of Klara. The floats, when released, all remain still. No physical or causal process need flow along the chains that rest, we may imagine, inert and with links not even in physical contact. Throughout this vast clockwork universe that we have appended to the armature an absolute stillness prevails; gears hang motionless in the air, not touching. But the supervenience thesis tells us that any mode of consciousness which occurs during the time of the computation must supervene on the physical activity and processes that occur through that period. Hence, Olympia's phenomenal state ϕ must supervene solely on the activity of the armature (and perhaps the floats), since that is all the physical activity there is. The masses of idle machinery form no part of the supervenience space, since nothing happens there. We have already seen, however, that, without the machinery, the armature alone cannot be interpreted as performing any nontrivial computation, and so, by the necessity condition, the conscious state ϕ cannot supervene on its activity alone.

In short, the computationalist is committed to the claims that the armature moving without the extra machinery hooked up cannot be conscious and that the system composed of the armature moving with the machinery hooked up must be conscious. But the physical activities that occur with and without the idle machinery connected are exactly identical, so these two claims contradict the supervenience thesis.

As has been already noted, the supervenience thesis implies that the presence or absence of inert, causally isolated objects cannot effect the presence or absence of phenomenal states associated with a system. In Olympia's case, the objects are huge, unwieldy, and unimaginably complex, but still inert, neither active in themselves not exerting any influence over the armature. So any existing phenomenal state does not supervene on them. But the supervenience space of Olympia's computational description, indeed whether she is computing at all, depends vitally on the counterfactuals that the idle machinery supports. Hence, her conscious phenomenal states cannot derive from her computational structure. A computational theory of consciousness is not possible.

Since a computationalist cannot give up the sufficiency condition, and since it would tear the guts out of the notion that mental states arise from a complex computational structure to give up the necessity condition, our central focus must be the plausibility and force of

the supervenience thesis. Here two avenues of approach are possible. Either the computationalist can argue that the thesis is fine but that, contrary to appearances, the physical activities with and without the machinery hooked up are not the same, or else he can try to renounce the thesis altogether. Let us consider these possibilities in turn.

The supervenience thesis depends upon some independent notion of physical activity, and, in our application of it, upon the notion of sameness of physical activity. It would be beyond our needs to try to formulate an exact definition of physical activity here, but one line of argument deserves our attention. Perhaps one might urge that the counterfactuals supported in the first case differ from those supported in the second case, and *ipso facto* the physical processes occurring in the two instances must differ. Now, the argument by *ipso facto* is pleasingly short and crisp, but carries little conviction. We do have intuitions about sameness and difference of physical activity, and we cannot allow the computationalist just to build whatever is handy into the individuating conditions for types of physical process. Otherwise, even the Cartesian dualist can embrace the supervenience thesis: wherever there is a difference in the state of a thinking substance, *ipso facto* there is a difference in physical activity. But the notion that difference of counterfactual structure implies difference in physical process does not fare well when confronted with cases.

Consider, for example, a deterministic pin-ball game.[8] If we set the ball off at a specified place and with a specified velocity, it will always trace exactly the same path through the board. The only physical processes involved are those associated with the motion of the ball down the board and the interactions of it with the pins that it encounters. If we now remove a few pins, pins which the ball never touches on this path, pins which perhaps it never even comes near, the ball will continue to retrace exactly the same path in exactly the same fashion. The physicist's explanation of why it traces just that path will remain exactly the same. No motion or energy transfer or change of state occurs anywhere outside the path of the ball.[9] From a physical point of view, the processes and actions that occur with or without the peripheral pins in place are identical, for there are no physical processes occurring outside the path. But the counterfactuals in the two cases are different. For had the ball been given a different initial push, a push that would have carried it into the region where the pins are removed, then its path would be different in the two cases. So, although the counterfactuals supervene on the entire physical state of the system, differences in counterfactuals about the evolution of the system need not imply differences in physical processes that are evolving at a time.

[8] I owe this illustration to Kevin Kelly.

[9] Of course, at a subatomic level, there are all sorts of motion, but that is not essential to the case. The facts would remain the same in a Democritean world where the matter is just solid and totally inert stuff.

To avoid getting bogged down in general claims, let us look more closely at the details of Olympia, and at the sorts of changes which are sufficient to defeat the support of counterfactuals provided by the blocked machinery. Above we considered the cases of Olympia running with and without the idle machinery present. Because of the immense quantity of machinery involved, one might misgive that its removal would necessitate some considerable change in the physical happenings associated with the machine. To alleviate such doubts, here are two cases in which the support can be neutralized by changes that can hardly be construed as altering the physical activities present.

An Argument by Addition: Suppose we run Olympia, fully connected, on τ so that (according to the supervenience thesis) the conscious state ϕ occurs. Now we reset her (and the tape) to run again, but we add a secondary block to each of the copies of Klara. The second block might be a thin piece of metal suspended between the frozen gear teeth. It need not be in physical contact with any part of the machinery. Now, however, were the first block to be removed (which will not, of course, happen when we run Olympia on τ from $S_{[0]}$ to $S_{[N]}$), the gears would contact the second block and jam. The copies of Klara no longer support the right counterfactuals, so on the second run Olympia is not conscious. But, given that the second blocks in fact never even touch any part of the machinery, exerting no physical influence or force at all, how could the physical activity taking place in Olympia during the first run be said to differ from that in the second? Speaking loosely, how could the rest of the system know that the blocks are even there?

An Argument by Subtraction: Suppose that, as we repeatedly run Olympia on τ, the chains connecting the floats to the blocks are slowly rusting. At first, associated with every run there exists a conscious state ϕ: Olympia feels, say, a toothache. Eventually, though, the chains so weaken that, were they to be pulled, they would break rather than unblock the machines. No unusual physical effect accompanies the passage of the point of critical weakness, nor does the progressive rusting cause any other alteration in the structure of the device. Still, according to the computationalist, once the critical point has been passed, suddenly no toothache accompanies the passage of the armature from left to right. Speaking loosely, how can the system know that the critical point has

been passed? How can the psycho-physical connection be broken by such a minor change in physical state which has no influence on the dynamical interactions occurring in the system?

The presence or absence of an inert second block that enters into absolutely no causal interaction with the system cannot plausibly be said to change the nature of the physical activities going on in the system. So the computationalist must renounce the supervenience thesis altogether. But even this does nothing to solve the puzzles presented above: even if mental states supervene on more than just the physical activity of a system, the crucial role of the entirely isolated block remains inexplicable. And, in countenancing the possibility of such effects, the computationalist would cut himself off from the research tradition from which the tradition grew. To see this, let us apply the point directly to brain activity.

The modern picture of brain function rests primarily on the notion of neural activity. The essential structure of mentation seems to be founded in patterns of neural firings. Because those firings can be analyzed as carrying information, the brain has come to be considered as an information processor. So let us suppose that some time in the future the electro-encephalograph is so perfected that it is capable of recording the firing of every single neuron in the brain. Suppose that researchers take two different surveys of a brain which match exactly: the very same neurons fire at exactly the same rate and in exactly the same pattern through a given period. They infer (as surely they should!) that the brain supported the same occurrent conscious state through the two periods. But the computationalist now must raise a doubt. Perhaps some synaptic connection has been severed in the interim. Not a synaptic connection of any of the neurons which actually fired during either period, or which was in any way involved in the activity recorded by the encephalograph. Still, such a change in connection will affect the counterfactuals true of the brain, and so can affect the subjective state of awareness. Indeed, the computationalist will have to maintain that perhaps the person in question was conscious through the first episode but not conscious at all through the second. I admit to a great degree of mystification about the connection between mind and body, but I see no reason to endorse such possibilities that directly contradict all that we do know about brain process and experience.

Whether the reason is enshrined in the supervenience thesis or not, our general picture of the relation between physical and mental reality firmly grounds the intuition that Olympia's experience cannot be changed by the presence or absence of the second set of blocks. These intuitions are not sacrosanct, but the computationalist especially abandons them at his own risk. For similar intuitions are often

appealed to in defending the appropriateness of computational analogies in the first place. One first step in arguments aimed at inducing assent to the possibility of computers that can think, or feel, or intend, is to imagine that some sort of prosthetie neuron made of silicon has been invented.[10] We are then to imagine slowly replacing some poor sap's brain bit by bit until at last we have a silicon brain that, our intuitions should inform us, can do all of the mental and intensional work of the original.

This is as yet a far cry from showing that anything has mental properties in virtue of its computational structure, but it is supposed to break down parochial species-chauvinistic views about there being any deep connection between mentality and organic chemistry. But the thought experiment rests on a tacit appeal to supervenience. How could it matter, one asks, whether the electrical impulses are carried by neurons or by doped silicon? The implication is that mentality supervenes only on the pattern of electrical or electrochemical activity. If the computationalist now must assert that the presence or absence of a piece of metal hanging untouched and inert in the midst of silent, frozen machinery can make the difference between being conscious and not, who knows what enormous changes in psychical state may result from replacing axons and dendrites with little copper wires? Should the computationalist reject the extremely general intuitions at play in assessing Olympia's case, no means of judging the plausibility or implausibility of any theory of mind seems to remain.

The silicon brain *Gedankenexperiment* is worthy of closer examination, as it turns on intuitions about supervenience similar to those we have invoked. But it requires a stronger supervenience thesis than ours, for the occurrent processes in the silicon brain are physically distinguishable from those in the organic brain in a way that the activities in Olympia with and without the extra blocks are not. Still, the suggestion that the pattern of electrical activity (however supported) determines the occurrent mental state has some plausibility, and allows us to abstract away from the organic features of the brain. Can we push the supervenience claim to an even higher level of abstraction?

The next obvious step is to suggest that the use of electricity (or electrochemical activity) is not essential to mentality. Let the ersatz neurons communicate by any causal process at all, and so long as the pattern of that activity is isomorphic to the brain, we still have consciousness and intensionality. At this stage of abstraction, we are

[10] Cf., for example, Pylyshyn's response to Searle, *op. cit.*, p. 442, or Clark Glymour's "Silicon Reflections" (typescript).

allowed to replace axons and dendrites with, e.g., pipes, and to let pulses of water play the functional role of electrochemical discharges. This allows a greatly broadened class of materials out of which a mind can be built. The appropriate supervenience thesis at this level of abstraction would assert that mental events supervene on the pattern of causal interaction in the brain, where the exact physical or chemical nature of the causal processes has been abstracted away. One neuron sending an electro-chemical impulse to another was first likened to one transistor sending an electrical pulse to another, and now to one tank of water sending an hydraulic pulse to another. Olympia, however, has not yet graced the ballroom, for she is structurally equivalent to the brain at none of these levels.

Olympia only becomes relevant at the next level of abstraction. From the generic two events connected by a causal process we now abstract off the requirement that any causal chain connect the two and posit only a counterfactual or information-theoretic connection. Now, no physical pulse of any description need pass between two objects for the relevant relation to hold. Our chains, lying slack and inert, support no physical activity at all, but they are nonetheless transmitting information. The intact chain transmits information to the blocked machine about the state of the tape, since, had the tape been different, the block would have been pulled. The rusted chain, lying equally slack and also with no causal process coursing through it, fails to transmit that information, since it would have broken if pulled. *Information transmission between two points does not require any physical activity or causal process connecting them.*[11] So two physical systems engaged in precisely the same physical activity can be processing information differently. And now Olympia can make her entrance.[12]

[11] Cf. Fred Dretske's *Knowledge and the Flow of Information* (Cambridge: MIT, 1981), pp. 26 ff.

[12] It may be useful, now that the argument is complete, to contrast this line of investigation with some other similar arguments. As has already been noted, I do not accept Searle's Chinese room case, since it does not establish that no mental state supervenes on the activity of the man in the room, only that any such state would be disjoint from that of the man in the room. The conclusion at which Olympia's tale arrives, however, is exactly that which Searle set out to prove: having a mind is not merely a matter of instantiating a program.

Arnold Zuboff, in "The Story of a Brain" [*The Mind's I*, D. Hofstadter and D. Dennett, eds. (New York: Basic Books, 1981), pp. 202–212], subjects a cerebrum to radical disintegration in order to undermine the notion that "what decisively controlled any particular experience of a man—controlled whether it existed and what it was like—was the state of his nervous system . . ." (p. 202). Zuboff first suggests that some of the electrical connections within the brain might be stretched, resulting in a system more spread out but still capable of the same patterns of neural activity. Then, in order to overcome time-lag problems that the stretching would introduce, the direct causal connections between neurons are cut completely, being replaced by "impulse cartridges" that stimulate the neurons at the same time that they would have been stimulated if the causal connection held. The collection of neurons is still supposed to display the same "pattern of activity" in this state. Finally, the individual causally isolated neurons are scattered across the universe. The claim that the same patterns of neural firing still support any experience is shown to have unacceptable consequences.

Zuboff's case poses problems only for a particular interpretation of what constitutes a "pattern of neural activity." This is not an interpretation that the computationalist would accept. Once the causal connections between neurons have been cut, the collection of neurons (even with the "impulse cartridges") would no longer instantiate the same program. For the program, as we have seen, specifies subjunctive connections between machine states, and hence between the elements that constitute the system. But these subjunctive connections are determined by the physical and causal links between the neurons. Originally, counterfactuals such as "Had neuron *A* not fired, neuron *B* would not have fired (although it did)" are made true by the direct causal link between *A* and *B*. Once the cartridges are introduced, *B* will fire independently of what *A* does, so the counterfactual will fail.

Indeed, for any functionalist view, the relevant identity of any given neuron in the system is determined by its functional–and hence causal–role in the system, not by some tag or name attached to the particular neuron. If I switch two neurons in your brain, then exhibiting the same pattern of neural activity is not having the same particular individual neurons fire in the old sequence. Rather, you exhibit the same pattern of activity only if each switched neuron fires when its counterpart used to fire. Each inherits the role of the counterpart in defining the pattern when it inherits the functional role of the counterpart. But, in Zuboff's disintegrated brain, there are no functional roles at all, and hence no "patterns of activity."

Far from causing the computationalist problems, Zuboff's case might be used to support a computational approach. Zuboff's notion of "pattern of neural activity" is clearly too weak to capture the important aspects of brain function. Computational structure provides a stronger criterion, one which defeats Zuboff's gambit, since computational structure is not preserved through the disintegration process. Olympia, however, shows that even this stronger criterion of identifying "patterns of neural activity" is too weak. We must take into account not only the subjunctive connections between parts of the brain but the active causal connections as well, for the "patterns" defined solely by program structure cannot guarantee experience.

The doleful tale of Olympia, then, does not purport to establish that things with minds must be wet and squishy. The silicon brain and the hydraulic brain may, for all we have said, be conscious. Nothing has been said for the supervenience theses associated with these levels of abstraction, but neither has any objection been brought against them. Olympia does show, however, that consciousness cannot supervene on the computational activity of the brain. Let us turn now to a careful consideration of exactly what else Olympia shows.

V. ASSESSING THE DAMAGE

Throughout our investigations so far, we have focused on necessary and sufficient conditions for consciousness. Consciousness is intrinsically fascinating, but it is also particularly well-suited to this line of inquiry, because it is an occurrent mental phenomenon. Toothaches and other subjective experiences are events—they take place and are completely realized at or through a particular period of time. This

fact strongly supports the thesis that they supervene on the physical events and processes that also occur through that time.

Other important mental properties, such as intelligence and understanding and intensionality, seem much more dispositional than consciousness does.[13] It is plausible to suppose that an individual's present intelligence depends on counterfactuals in a way that his present tickles and itches do not. The more dispositional a property appears, the easier it is to contend that it supervenes not only on the activities but also on the dispositions of a system. And, since the dispositions of Olympia are changed by the presence or absence of the second block, Olympia may not threaten a computational account of these other properties. The computationalist may then drop consciousness from his list of potential quarry—such a spooky phenomenon is usually not high on the list anyway—and be content to occupy himself with accounts of intelligence, understanding, intensionality, and the like. Perhaps his computers will never experience anything, but still someday they will think and speak with meaning. These goals seem quite magnificent enough.

I do not think that the outlook for the computationalist is quite so rosy. Although intelligence is more a capacity than an occurrent phenomenon, still capacities are realized in particular acts and activities. So, if having a certain program is sufficient for being intelligent, then solving a particular problem by operating according to the program is sufficient for solving the problem intelligently, and, presumably, running some nontrivial program is necessary for solving it intelligently. So let us build a version of Olympia in which the movement of the armature will instantiate the bit of that program used in solving that particular problem. (Recall that, for any given program and any finite sequence of operating that program on a given input, we can build Olympia to have that program and to operate that sequence just by the motion of the armature.) When the armature moves with the secondary blocks not in place, Olympia solves the problem intelligently, but, if we move the blocks over a few centimeters and repeat the process, she will solve the problem stupidly. Or let the output of the program be interpreted as linguistic behavior, the program designed to impart linguistic understanding. Blocks

[13] There is a sense, I think quite legitimate, in which intelligence, understanding, and intensionality all presuppose consciousness, and hence according to which the failure of computational theories of consciousness would automatically rule out computational accounts of any of these. Since most of the artificial intelligence community maintains a studious silence on questions of consciousness, I infer that they do not use these terms in this sense. The following discussion is directed toward this other sense, the sense, for example, of "belief about the ambient temperature" according to which a thermometer can have such a belief. I do not pretend to understand this other sense.

out, and Olympia speaks with true intensionality and insight; blocks in, and she produces the very same strings in the very same way but does not mean anything by it.[14]

Ultimately, Olympia has demonstrated that computationalism, which first seemed to be a version of functionalism, actually passes to a level of abstraction which undercuts functionalism's central insight. The functionalist demands that we look beyond the surface behavior to the causal structure of the processes that produce it. A system must not only get the output right; the causal structure responsible for the output must also be of the right form. When Olympia runs π on τ, whether in writing a sentence or solving a problem, all of the blocked machinery, including the presence or absence of the secondary block, plays absolutely no causal role in producing that behavior. So, for the functionalist, the presence or absence of the block, or of the machinery as a whole, can make no difference to the intelligence, intensionality, or consciousness of the system during that computation. But, for a computationalist, the detailed structure of these inert and silent jungles of machinery makes all the difference. The verdict of the causal role functionalist is immensely the more plausible.

In *The Sandman,* when Nathanael discovers Olympia's true nature, he goes mad, finally casting himself down from a tower in a fit of violent rage.[15] We need not counsel the computationalist to anything so drastic. Our Olympia demonstrates that running a particular program cannot be a sufficient condition for having any form of mentality. But computational structure may yet be a necessary condition for consciousness or intelligence. Certainly, nothing in these arguments suggests that consideration of counterfactuals or dispositions is irrelevant to questions of mentality and consciousness. And one way, albeit not the only way, of specifying some of the dispositions of an object is by providing a description of its computational structure. If a system must be describable as running a certain kind

[14] This form of argument should sound familiar to the AI enthusiast, for it is an echo of the very arguments AI proponents use. Witness Pylyshyn's deployment of the silicon brain argument which concludes: "Thus if more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output *function* of each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now except that [according to Searle] you would eventually stop *meaning* anything by it" (Searle, *op. cit.,* p. 442). I leave the reader to consult her or his own judgment about the strength of the intuition that replacing protoplasm with silicon cannot rob one of intensionality compared with the intuition that putting the secondary blocks in cannot rob Olympia of hers.

[15] Actually, Nathanael's final insanity is brought on when he discovers something (but what?) about the nature of Klara. I leave you, gentle reader, to consider the implications of this remarkable fact.

of program in order to be intelligent, however, it is a further constraint that the counterfactual relations that constitute the program structure must be supported directly by the nature of physical activity in the system. Computational and causal structure are intimately intertwined in brain activity, a fact that any acceptable theory of mind must reflect.

The segregation of causally-active from counterfactual–supporting structure in Olympia is bought at a great price. In order to minimize the activity involved in running π on τ, Olympia had to be bloated with a vast galaxy of apparently redundant machinery. Indeed, one way to understand Olympia's structure is by analogy: Olympia creates the illusion of being Klara in action in much the same way that a movie creates the illusion of motion on the screen. Like still frames of celluloid, each blocked copy of Klara has captured in an inert object a moment in the career of her operation. And as the projector's light brings momentarily to life each still photo in sequence, so the movement of the armature, the successive unlatching of the floats, momentarily makes the whole dispositional structure of each copy of Klara the dispositional structure of Olympia herself. The problem with computationalism is that it does not contain the conceptual resources to distinguish the flickering illusion from reality.

Only a philosopher would consider instantiating a program in such a perverse way as in Olympia. Perhaps the pragmatic constraints that govern real AI research will naturally drive workers to create instantiated programs with just the right blending of computational and causal structure to support consciousness. But as yet we have no proof that pragmatic constraints should have any tendency to lead to this fortuitous result. At present we know neither in what direction the demands of working with digital computers exert pressure nor in what direction true artificial intelligence lies.

We must consider carefully whether the computational level of abstraction plays any part in defining consciousness and intelligence, and, if it does, what other level of description must also be invoked. Perhaps we must speak of electro-chemical structure, perhaps only more generically of the patterns of causal process in a system. Olympia has shown us at least that some other level beside the computational must be sought. But, until we have found that level and until we have explicated the relationship between it and computational structure, the belief that pursuit of the pure computationalist program will ever lead to the creation of artificial minds, or to the understanding of natural ones, remains only a pious hope.

TIM MAUDLIN

Rutgers University