

Article

Against Functionalist Theories of Consciousness

MICHAEL V. ANTONY

By far the most prevalent contemporary philosophical theory of consciousness is functionalism. According to one common and simple formulation, conscious states and processes are conscious rather than not in virtue of their particular causal relations to inputs, outputs, and other mental states and processes. Such a view has much to recommend it. When coupled with a token identity theory, it exhibits the dual virtues of respecting materialism and allowing for alternative physical realizations. And in its computationalist guise, it accords favourably with the rapidly expanding cognitive scientific and neuroscientific enterprises.

Such merits notwithstanding, many have remained unconvinced. A number of well-known arguments have been marshalled against functionalist theories of consciousness, generating a vast literature. Most aim to show, in effect, that two functionally identical systems could nonetheless differ in regard to their conscious states and processes.¹ In this paper I take an opposing tack and argue that with respect to any functionalist theory whatever, there could be one system that realizes it and another that does not with no difference in consciousness across the two systems.

This paper derives from an essay in Antony (1990). I am indebted to Ned Block, Martin Davies, Joe Levine, Georges Rey, Bob Stalnaker, Eddy Zemach, an anonymous reviewer for *Mind and Language*, and to audiences at The Hebrew University of Jerusalem and Ben-Gurion University of the Negev, for helpful comments on various drafts. Work on this article was supported in part by the Social Sciences and Humanities Research Council of Canada, and by an Israeli Vatat Fellowship. Address for correspondence: Department of Philosophy, Tel Aviv University, Tel Aviv 69978, Israel.
Email: antony@ccsg.tau.ac.il.

¹ The inverted spectrum and absent qualia arguments are perhaps the best known of this class. On inverted spectra and absent qualia see e.g. Shoemaker, 1981b, and Block, 1978, respectively.

I call the argument I develop the 'System Argument'. Like more familiar anti-functionalist arguments with respect to consciousness, the System Argument depends crucially upon pre-theoretical intuitions about consciousness. However, the intuition driving the argument is notably different from intuitions employed in the standard arguments.²

The System Argument is presented in Section 1, and the remainder of the essay is taken up with various matters of clarification. In Sections 2 and 4, I examine two arguments that are superficially similar to the System Argument in order to illuminate the System Argument's distinctive features and help ward off possible confusion. In Section 3 I provide a more refined characterization of the intuition underlying the argument. In Section 5 a natural objection to the System Argument is considered and subsequently put aside. A brief conclusion is offered in Section 6.

1. *The System Argument*

Imagine an individual, Sam, who begins to have a conscious experience at some time t_1 .³ Any conscious experience will do, as will become evident, but for tradition's sake suppose it is a pain. According to functionalism, roughly, Sam is in pain in virtue of there being some token physical state or process in his brain that is suitably related to other token physical states or processes in his brain, as well as to inputs and outputs (e.g. environmental stimuli and bodily motions).⁴ Since the relations commonly

² Tim Maudlin (1989) has advanced an argument that rests on an intuition similar to that of the System Argument, and involves many of the same issues. Maudlin's argument differs from the System Argument in numerous respects as well, however, perhaps the most important of which is that while his argument applies only to computational theories of consciousness, the System Argument applies more generally to all functionalist accounts of consciousness. I leave a more thorough discussion of Maudlin's argument for another time.

³ I assume in this paper that the temporal boundaries of conscious experiences can be fixed at least as accurately as can the temporal boundaries of the neurological states and processes that realize them. In Dennett's (1991) book *Consciousness Explained*, however, he argues that the temporal boundaries of experiences cannot be determined with precision; that there are no facts of the matter as to when exactly experiences begin or end. (See also Dennett and Kinsbourne, 1992.) I am unpersuaded by Dennett's arguments, e.g. for many of the reasons developed in Block, 1993. However, even if Dennett turns out to be right, since his conclusions entail that functionalism about consciousness is false (since every neurologically realized functional state or process has determinate temporal boundaries) (see Antony, 1992), at least the System Argument's conclusion will be true, if not all of its premises.

⁴ More accurately, the interrelated tokens in Sam's brain realize a functionalist theory in virtue of instantiating certain appropriately related first-order (e.g. neurophysiological) properties. (For details see e.g. Block, 1978, 1980b; Lewis, 1972; Loar, 1981; Shoemaker, 1981a.) That said, it often will be convenient to speak of the causal relations of *token* brain states and processes, rather than those of the first-order properties in virtue of which the tokens bear such relations.

appealed to in functionalist theories of mind are causal relations, we can say that Sam's token brain state or process is a pain on such theories because it is of a type that occupies the appropriate *causal role* for pain in the operation of his brain.

Let us call the part of Sam's brain where his pain is realized at t_1 his 'pain region',⁵ and suppose it is neural activity in that region that realizes his pain at t_1 . Imagine further that given the overall state of Sam's brain at t_1 (i.e. given all inputs and other currently realized states and processes), the activity in his pain region is disposed to cause activity in some distinct region of his brain—his 'action-planning region'—where plans are constructed for actions like the fetching of an aspirin, for example. Lastly, assume that it is in part *because* the activity in Sam's pain region is causally related to future activity in his action-planning region (given his overall brain state) that the former activity is pain, according to functionalism.⁶ Assume, in other words, that all it is for a token state or process to be pain, according to functionalism, is for it to have the very sorts of causal relations had by the activity in Sam's pain region at t_1 .

At t_1 , then, Sam begins to experience a pain, and some time later, the construction of a plan to fetch an aspirin, say, begins. This sequence of events is realized, first, by activity in Sam's pain region, starting at t_1 , and then activity in his action-planning region beginning some time later. The latter activity is later because it is caused by the former activity. We might suppose, for example, that the causal interaction between the activity in the two regions occurs via neural pathways that connect the two regions, and that it takes roughly 100 milliseconds for the signal to reach the action-planning region from the pain region.

Now for a piece of science fiction. Imagine that there exists some impressive technology by means of which one can instantly incapacitate (or destroy) any region of an individual's brain one chooses, right down to components of individual cells. That done, imagine once again that activity in Sam's pain region has begun at t_1 , and that the signal is travelling toward the action-planning region, but thus far has reached only the half-way point along the connecting neural pathways. Call the time at which the signal reaches the half-way mark ' t_2 '. Now assume that Sam's pain region has been activated since t_1 , and that it remains activated at t_2 . Suppose, finally, that at t_2 the action-planning region of Sam's brain is incapacitated with our imaginary technological wonder. Two questions arise.

⁵ I call the region his 'pain region' for ease of reference, but it would more accurately be called his 'region where certain pains are realized at and around t_1 '. The point to be emphasized is that the argument does not assume that there is a single region of Sam's brain where all pains occur, or that there are brain regions that subserve only pain, but only that each pain occurs some place or other. Similar points holds for the 'action-planning region' to be discussed presently.

⁶ If the reader doubts that any plausible functionalist account of pain would refer to action-plan constructions, any alternative effect may be substituted.

First: according to functionalism, is Sam experiencing pain at t_1 ? When Sam's pain region first becomes activated at t_1 , his action-planning region is as yet untouched. At t_1 , therefore, Sam's brain is fully intact and is in perfect working order. In particular, all the underlying physical conditions necessary for pain-region activity to cause action-planning region activity (and for it to cause other effects, given different inputs and concurrent states, etc.) are in place in Sam's brain at t_1 . Therefore, since the activity in Sam's pain region at t_1 bears all the requisite causal relations for pain within the system of Sam's brain, according to functionalism, that token activity is *pain* on that view.

Second: according to functionalism, is Sam experiencing pain at t_2 ? Notice first that because Sam's pain region has remained activated from t_1 , Sam is in pain until t_2 for the same reason he is in pain at t_1 : until t_2 , when his action-planning region is incapacitated, Sam's brain is unimpaired and is operating flawlessly. The story changes at t_2 , however, once Sam's action-planning region is incapacitated. For at t_2 the activity in his pain region no longer bears the necessary causal relations for pain, according to functionalism, since it is no longer disposed to cause an action-planning process due to the incapacitated region of his brain. The physical conditions necessary to underwrite such a disposition are simply not present at t_2 in the (now modified) system of Sam's brain. Accordingly, the activity at t_2 in the part of Sam's brain we have called his 'pain region' is *not a pain*, by functionalist principles.⁷

⁷ One might argue that incapacitating the action-planning region at t_2 does not eliminate the *disposition* of pain-region activity to cause action-planning-region activity, but merely prevents the occurrence of some causal interaction. For the former activity is still disposed to cause the latter activity, *given that the action-planning region is intact*. The following counterfactual holds, for example. If Sam's action-planning region had not been incapacitated, pain-region activity would have caused action-planning-region activity. Thus, one might conclude, the activity in Sam's pain region at t_2 is pain according to functionalism. The argument, however, rests on a confusion. Consider two dispositions: (1) the disposition of pain-region activity to cause action-planning-region activity, given certain inputs and other currently realized states and processes; (2) the disposition of the former activity to cause the latter, given those same inputs, states and processes, *and* given that there are intact system parts to realize the action-plan-construction (and the inputs, states and processes). It is this second disposition the critic rightly claims is left untouched by the incapacitation. However, only the *first* disposition is relevant to the functionalist account of pain, and that disposition clearly is eliminated. Functional theories, after all, specify causal relations among inputs, outputs, and mental states and processes, plain and simple; they do not state conditions that there be intact system parts to realize such inputs, outputs, states and processes. Moreover, if they did, it would seem, anything could realize anything. For example, the temperature of my coffee cup could realize pain, since it is disposed to cause an action-plan-construction, given the inputs and concurrent states and processes required by the theory, and given that there are appropriate system parts to realize the plan-construction and all else that needs realizing. But of course that will not do. Dispositions of the second sort, therefore, are of no use to functionalism.

Functionalism thus dictates that Sam is in pain from t_1 until just before t_2 , but not at t_2 . But notice what has produced the change at t_2 : the incapacitation of an entirely *inactive* brain region. At t_2 , recall, no signal from Sam's pain region has yet reached his action-planning region. The signal that left his pain region at t_1 , for example, has reached only half way across the connecting pathways by t_2 ; and the signals that left after t_1 are further away still. Prior to the incapacitation at t_2 , therefore, Sam's action-planning region is sitting utterly dormant; it is inert brain matter waiting to be engaged.

But now how could tampering with an *idle* region of an individual's brain cause a change in that individual's conscious experience *at the time of the tampering*? Surely, it would seem, the state of one's unused brain parts at a time is no more relevant to the character of one's experience *at that time* than is the state of the moon. However functionalism implies otherwise, as we have seen. It requires that Sam is in pain from t_1 until the incapacitation at t_2 , after which time his pain ceases. That is an extraordinary result, but, I submit, wholly implausible. Fiddling with idle brain parts, one would have thought, cannot affect the nature of one's experience. The functionalist account of pain assumed at the outset, therefore, would appear to be false.

Notice that it was entirely arbitrary in the presentation of the argument that the experience appealed to was a pain, and that the effect was the construction of a plan to get an aspirin. Any other experience would have done as well—an itch, an auditory image, etc.—as well as any other effect. Notice, moreover, that it is immaterial that the selected experience was a state rather than a process.⁸ And similarly with the effect: although the construction of a plan is perhaps best seen as a process, other examples

⁸ Actually, running the argument on functionally defined processes requires that a somewhat more complex story be told. Processes essentially involve change, and types of processes, it would seem, must be identified in part by the types of changes that occur during them. Processes, in other words, must have an *internal structure* of some sort. This suggests that a functionalist account of processes cannot be given solely in terms of a process's causal role. For in principle there could be tokens with the same causal role as some process P , but which lack P 's internal structure, or have none at all. Such tokens, intuitively, would either not be tokens of P , or not be processes. Now I do not wish to enter into how functionalism ought best to account for the notion of a process. Any account, presumably, will involve something like a sequence of functionally defined states. A pain process, for example, might be defined as a sequence consisting of states S_1 – S_5 , where each state is defined in the usual way by its causal role. (See Rey, 1993, for a theory of sensations along these lines.) My present concern is just to point out that such an account poses no threat to the System Argument, for the argument simply can be run on the individual components, S_1 – S_5 . Consider, for example, a system that enters S_5 , and immediately thereafter has one of its inactive parts incapacitated—a part needed for realizing one of S_5 's required effects. S_5 would no longer be instantiated after the incapacitation, according to functionalism. But the nature of the experience, so the argument goes, would be unaffected. Therefore, whether S_5 is realized is irrelevant to whether the pain is. And *a fortiori* with respect to the entire sequence of states, S_1 – S_5 .

more akin to a state would have sufficed as readily. Exactly the same argument, therefore, could have been advanced with *any* conscious state or process, as well as with any effect.

The reason the argument generalizes over all types of conscious states and processes, and all types of effects, is that the argument proceeds at the level of the *realizations* of the states or processes; and any realizations will do to make the point. Indeed, all that is required is that the experience be realized *before* its effect—and that is guaranteed by the fact that (typically at least) causes precede their effects. That given, once the experience has been instantiated, it will always be possible as a matter of principle, before the effect would normally be realized, to make it so that the effect cannot be realized. That can be done by tampering with the idle system parts that would be used in realizing the effect a short time later. (Or the parts might be imagined to be damaged coincidentally by some natural process at the appropriate time.) But now if it is assumed that what realizes the experience remains instantiated when the tampering occurs, functionalism entails that the experience must cease at the time of the tampering; and that simply is not plausible.⁹

Any experience and any effect, therefore, will suffice for the presentation of the System Argument. Accordingly, the above considerations extend well beyond the particular case of Sam's pain and amount to a quite general argument against functionalist theories of consciousness.

2. *The Blowup Argument and Functionalism's Causal Relations*

Here is an argument that is similar to the System Argument. Suppose that Sam is in pain at t_1 . And suppose that in accordance with the functionalist account of pain, and given Sam's other currently realized mental states and processes, that pain is disposed to cause the construction of a plan to fetch an aspirin. Suppose, moreover, that due to the natures of the parts of Sam's brain that realize the pain and the construction of the plan,

⁹ In setting up the argument, Sam's action-planning region was assumed to be distinct from his pain region. It should now be evident that the argument needs that assumption to ensure that Sam's action-planning region can be idle while his pain region remains activated. For the same reason it must be assumed that his action-planning region subsumes no other function. One might object, however, that such assumptions cannot freely be made since pains, plan-constructions, and other functional kinds, might be realized in the *same hardware*, e.g. in distinct patterns of activity over the same set of neurons. Well of course that is possible, but it is not necessary. Functionalism explicitly allows for alternative ways of realizing the same theory, and the above argument exploits one such way. If functionalism fails with respect to that realization of that theory, therefore, functionalism fails with respect to that theory. The one way out would be to restrict realizations of functionalist theories to those in which the realizations of functional kinds share hardware, a move that would be both implausible and *ad hoc*.

the plan-construction cannot be effected until 100 milliseconds after the pain has first been realized. So far all of this is like the System Argument. But now what if Sam's action-planning region is incapacitated a short time after t_1 , before the aspirin-plan-construction can get under way? Or, more dramatically, what if the Earth *blows up* shortly after t_1 ? Surely Sam would still be in pain from t_1 until the incapacitation or explosion (assuming his pain region remains active until then), even though the pain does not bring about its required effect, according to functionalism. Since functionalism implies otherwise, however, functionalism is false. Call this the 'Blowup Argument'.

The Blowup Argument bears a striking resemblance to the System Argument, and consequently the two can easily be confused. However, the Blowup Argument is a bad argument. In order, therefore, that criticism appropriate to the Blowup Argument not be misdirected at the System Argument, it is important that the differences between the two arguments be clearly seen. That is one reason for discussing the Blowup Argument. A second is that comparing the two arguments illuminates the conception of *causal relations* that the System Argument assumes is part and parcel of functionalism; and as a result the System Argument is brought into tighter focus.

So what is so bad about the Blowup Argument? The problem is that it gets functionalism wrong. It assumes falsely that functionalism requires of token mental states and processes that they *actually bring about their effects* in order to be the kinds of states or processes they are. As if it were something functionalism denies, the Blowup Argument insists that Sam is in pain from t_1 until the explosion (or incapacitation) even though the activity in his pain region never ends up causing activity in his action-planning region. But of course, Sam is in pain until the explosion because his pain-region activity bears the appropriate causal relations within the system of his brain to count as pain. After all, Sam's brain is fully intact and is functioning perfectly right up until the explosion, so there is no reason why his pre-explosion mental states should be affected. This is just like a calculator that gets blown up in the midst of an arithmetical computation: in spite of the explosion, it is still true that the calculator was computing up until it was destroyed.

All this suggests that functionalism does *not* require of token mental states and processes that they actually bring about their effects, contrary to what the Blowup Argument assumes. Not requiring that makes sense. Functionalist theories are realized by *systems*, and systems inevitably cease to exist. Living creatures with minds, after all, die. And robots whose brains are destroyed lose their minds. Now, in principle, a system with a mind can be destroyed while realizing any (or almost any) mental state or process whatever. A system can meet its end while accessing a memory, shifting its visual attention, experiencing a chill, etc. Given that, however, it follows that any plausible functionalism must allow, at least with respect to most mental states and processes, that they can be realized even though

none of their prescribed effects are ever brought about. The sort of causal relation that one token must bear to another in order to count as a realization of a functional kind, therefore, cannot, in general, be *actual causation*.¹⁰

But then in virtue of what does a token physical state, process, etc. count as a realization of a functional kind? I would suggest that it is this: *the token must be properly situated in a system of the right sort*. For our purposes, a *system* can be conceived of, roughly, as a set of physical conditions that allow for specific sorts of causal interactions among tokens. A system is of the *right sort* if the specific types of causal interactions it supports map appropriately onto the set of causal relations dictated by the functionalist theory in question. And a token is *properly situated* in the system for it to be an instance of a functional type F if it gets paired with F in the aforementioned mapping (in virtue of instantiating an appropriate first-order property). Now so long as a system remains intact, a token properly situated in it can realize a functional kind F, even if that token never causes any of its prescribed effects, as a result of the system getting destroyed, for example. On this view, therefore, it is *system integrity* at time *t* that is relevant to whether a functional kind F is realized at *t*, not whether tokens at *t* actually bring about F's prescribed effects. It is this view of functionalism's causal relations that the System Argument assumes.

We are now in a position to contrast the Blowup Argument with the System Argument. As we have seen, the Blowup Argument takes functionalism to deny that Sam is in pain before the explosion, and that is where the argument goes wrong. The System Argument, on the other hand, assumes precisely the opposite, that according to functionalism Sam is in pain before the explosion. That is because Sam's brain is intact at *t*₁, and the activity in his pain region is properly situated in his brain to

¹⁰ It is important to note that the counterfactual nature of functional states and processes is not being ignored in the above discussion, but rather assumed. In speaking of a state's 'prescribed' or 'required' effects according to its functionalist account, I mean the state's required effects given the system's current inputs and other currently realized states and processes. (I shall also speak, interchangeably, of a functional kind's 'proprietary' causal relations.) Certainly the prescribed effects of states and processes will differ under various counterfactual circumstances, when there are different inputs and concurrent states and processes. The point of the above paragraph, accordingly, is not that states and processes need not, under various *counterfactual* circumstances, bring about the effects that are actually required of them; it is that the effects that a state or process's functionalist account prescribes for it in the actual set of circumstances need not be realized *actually*. In effect, this amounts to the claim that inputs to a system that are damage-causing (e.g. an explosion) are not included in the functional definitions of the system's states and processes. Once again a calculator illustrates the point well. The only inputs relevant to the identities of a calculator's computational states are those from the keypad; explosions, impacts with hard surfaces, and the like are excluded from its states' definitions. (This point is closely related to that made in n. 7.)

count as pain. What the System Argument takes functionalism to deny is that Sam is in pain *after* the incapacitation at t_2 —and that, even though the incapacitation consists wholly of tampering with an unused, inactive brain region.¹¹ Functionalism must deny that Sam is in pain at t_2 not because the activity in his pain region at t_2 will not bring about its proper effects—even though it will not—but because the system of Sam's brain at t_2 has been altered, and consequently the activity in his pain region at that time does not bear the appropriate causal relations within the newly altered system to count as pain. The pain-region activity is no longer disposed to bring about the required effects since it is no longer in a system of the *right sort*.

The System Argument proceeds by showing that the nature of a system's conscious experience can be changed, according to functionalism, by altering the features of the system in which that token experience is situated in a way that, intuitively, should have no effect on the system's consciousness. It is for this reason I call the argument the 'System Argument'.

3. Clarifying the Intuition

Sam is in pain, according to the System Argument, not only before the incapacitation of his action-planning region at t_2 , but also after it. Functionalism, however, must deny that Sam is in pain after t_2 . What supports the claim that Sam is in pain after the incapacitation is, at bottom, an *intuition*—one to the effect that manipulating a system's inactive or unused parts can have no effect on that system's experience at the time of the manipulation. When the incapacitation occurs at t_2 , recall, Sam's pain region is still active, and the signal that left that region at t_1 is only half way to the action-planning region. Sam's action-planning region at t_2 gets described in a number of ways—as inactive, unused, dormant, inert, waiting to be engaged, idle. But *which* property of his action-planning region, exactly, makes it such that manipulating it can have no effect on his experience? The System Argument, as presented above, provides no clear answer. Most of the notions expressed are related to inactivity. But while describing Sam's action-planning region at t_2 as inactive, or idle, or dormant suffices, I believe, to secure the argument's conclusion in its above form, the property that is operative in the argument needs to be characterized more abstractly for the System Argument to apply generally across all functionalist realizations.

The trouble is that putting the point in terms of inactivity—or worse, the absence of *neural* activity—restricts the System Argument's application to systems in which the realizations of conscious states and processes

¹¹ Notice that because the System Argument maintains that Sam is in pain after the incapacitation, it cannot even be run with an explosion.

involve (neural) activity. True, neural activity is commonly taken to realize the mental states and processes of humans. However, activity, neural or otherwise, need not be what realizes the states and processes of other types of systems, like robots or Martians. Token pains in Martian brains might be realized by parts that are blue, or by various other property instantiations. Moreover, even in the case of humans, it is accidental (if indeed true) that what realizes mental states and processes is neural activity. It might have been (or indeed might be) structural properties caused by neural activity, for example, that count. What is needed, then, is a way of expressing the System Argument's underlying intuition that allows the argument to generalize over the contingencies of particular types of realizations.

I think a way is close at hand. In the presentation of the System Argument, neural activity was assumed to realize Sam's mental states and processes, and the type of brain-part imagined to be incapacitated at t_2 without consequence to Sam's experience was an inactive part. Such a part, therefore, *was not involved in the realization of any mental states or processes in the system of Sam's brain at t_2* . For otherwise it would have involved activity. Now that, I submit, is the crucial property of Sam's action-planning region at t_2 that drives the argument—the fact that it is not realizing any mental states or processes; the fact that it is *unused*. Putting the matter this way allows the System Argument to generalize across all possible realizations. If, for example, in some particular system, all mental states and processes, conscious or not, are realized by parts with a certain structure, incapacitating parts that lack that structure cannot affect the nature of the system's experience at the time of the incapacitation. And similarly for all other types of realizations.

4. *The Paralysis Argument*

Consider now a second argument that closely resembles the System Argument. As with the Blowup Argument, it is worth examining so as to avoid confusing it with the System Argument, as well as to highlight some of the System Argument's distinguishing features.

A standard way of arguing against philosophical behaviorism (roughly, the view that mental states, processes, etc. are dispositions to behave) is to imagine a completely paralyzed person, e.g. a quadriplegic, and claim that contra behaviorism there is every reason to suppose that such a person could be the subject of mental states and processes. Call this the 'Paralysis Argument'. Such a person could be in pain, for example, even though all or most of his behavioral dispositions are lacking (e.g. withdrawing a finger when it contacts a hot stove). Now the Paralysis Argument is equally effective against functionalism, provided that the same mental-behavioral connections appealed to by the behaviorist are specified in the functionalist's definitions. Indeed, it might seem that one could go further

and take the Paralysis Argument completely 'inside the head', applying it to causal relations among internal states and processes specified by the functionalist. With respect to an account of pain that involves a causal relation to the construction of a plan to get an aspirin, for example, it might be argued that one could be in pain even if the part of one's brain that realizes the plan-construction were paralyzed or incapacitated. At which point the Paralysis Argument begins to look rather like the System Argument.

In fact, this internalized version of the Paralysis Argument differs from the System Argument in at least three ways. First, notice that the anti-behaviorist version of the Paralysis Argument proceeds by considering mental-behavioral connections between *specific types* of mental states or processes and behaviors (e.g. between a pain and saying 'ouch'), and then concluding that such connections are inessential to the identities of the states or processes under consideration. Such conclusions are grounded in actual or imagined cases of paralysis in which we are willing to attribute a type of mental state or process in the absence of some or all of its purportedly constitutive dispositions. The internalized version of the argument, therefore, insofar as it works, must proceed in the same fashion. Causal relations between specific types of states or processes must be considered, and argued to be inessential by imagining cases of 'internal paralysis' in which, intuitively, one state or process is realized in the absence of one or more of the causal relations specified by its functionalist account. In contrast, the System Argument, as we have seen, is such that the types of states and processes chosen for illustration are wholly irrelevant to the argument. It matters not in the least, for example, that pain and the process of constructing a plan to get an aspirin were chosen; any others would have done equally well.

A second difference between the two arguments is that, unlike the Paralysis Argument, the System Argument does not simply begin with an imagined case of paralysis and then claim that the mental state under consideration is realized nonetheless. Rather, it begins with the causal relations that the functionalist theory specifies *intact*, and shows how a system that is realizing a conscious experience at a given time can be *changed* to a system that is not—merely by tampering with currently unused parts of that system. Putting the same point another way, it is irrelevant to the Paralysis Argument, though not to the Systems Argument, whether the paralyzed brain regions are in use at the time of the paralysis. This irrelevance is explained by the first difference. Since the Paralysis Argument aims to show that causal connections between states or processes are inessential to their identities in virtue of the types of states or processes they are, it does not matter when the paralysis occurs so long as the disturbed connections are among states or processes of the right sort. With respect to the System Argument, on the other hand, the types of states or processes employed are unimportant so long as the causal relation between them is disturbed at the right time—when the brain regions to be incapacitated are not in use.

The third difference between the arguments is this: the Paralysis Argument is not directed at conscious mental states and processes in particular, but applies equally well to states and processes that do not involve consciousness. The opposite, however, seems to be true of the System Argument: that incapacitated brain parts are *unused* appears to have no force against functional characterizations of non-conscious states or processes.

To see this, we must imagine a case of a functionally defined non-conscious state or process, and an effect it is disposed to cause, in which, prior to considering the System Argument, intuitively the state or process cannot be realized in the absence of the disposition to bring about the effect. In other words, we need a case in which the Paralysis Argument is ineffective.¹² That is because the System Argument exploits a special case of 'paralysis'—one that involves incapacitating unused system parts. As a result, wherever both the System and Paralysis Arguments are effective, it will be unclear whether the System Argument's efficacy is due to the fact that *any* paralysis would have sufficed, or to the fact that the incapacitated parts were unused. Such cases, therefore, are inconclusive regarding the System Argument's effectiveness. The proper testing ground for non-conscious states and processes, accordingly, is where simple paralysis fails: if the System Argument is effective under such conditions, we can conclude that its appeal to unused system parts has force; if it is ineffective, we can be confident that cases in which the argument does work have nothing to do with the argument's distinctive features, but only with the fact that paralysis was involved—in which case we can conclude that the System Argument as such is ineffective with respect to non-conscious states and processes.¹³

Consider, then, some functionally characterized non-conscious state or process, and some effect to which, intuitively, the state or process must

¹² There are many cases in which the Paralysis Argument carries no weight; namely, wherever intuitions are lacking that a particular state or process could be realized in the absence of one or more of its proprietary causal relations. It is no argument against functionalism, for instance, simply to say that one could be in pain even if the parts of one's brain involved in thinking about one's pain were paralyzed; or that there could be a non-occurrent belief even if its connections to desire-like states and action-planning capacities were paralyzed. For the Paralysis Argument to have force (as we have seen), it must appeal to an intuition that the specific type of state or process under consideration can be realized in the absence of the relevant causal relation. However, the above cases, if they involve intuitions at all, involve intuitions that *favor the functionalist*, and so cannot be exploited in a Paralysis Argument against functionalism.

¹³ The same reasoning applies to the question of how we can be sure the System Argument's distinctive features are operational with respect to *conscious* states and processes. However, I have attempted to argue that the System Argument is effective in *all* such cases, no matter which conscious states or processes and which effects one chooses. *A fortiori* it is effective wherever the Paralysis Argument is not. Hence we can conclude that the System Argument's appeal to the incapacitation of unused brain parts is operative in the case of consciousness.

be causally related (given current inputs, etc.) to be the kind of state or process it is. An example might be a non-occurrent belief which, given currently realized non-occurrent desires (and perhaps other states or processes), is disposed to cause the construction of a plan for some action. Any state coinstantiated with such desires, and causally unrelated to any capacity for action-planning, it might be thought, simply would not be a non-conscious belief. (If this example does not work for the reader, one that does should be substituted.) In this case, then, the Paralysis Argument is ineffective. But now it seems clear that if we run the System Argument on this example, and eliminate the required relation by incapacitating the parts required for realizing the action-plan-construction *specifically when they are not in use*, we buy nothing over and above what we obtain with the Paralysis Argument: in both cases, intuitively, it seems the belief is no longer realized, just as the functionalist account dictates. With respect to non-conscious states and processes, therefore, it appears that the System Argument works if and only if the Paralysis Argument does. Incapacitating system parts that are unused, evidently, is ineffective with respect to nonconscious states and processes.

5. An Objection

I now wish to address an obvious and natural objection to the System Argument. In effect, the objection amounts to the claim that the System Argument *misrepresents* functionalism; and that once functionalism is properly characterized, it can easily accommodate the intuition the System Argument generates. Since the objection can take a number of forms, depending on the version of functionalism the objector favours, it is perhaps best seen as a class of objections. Each variant, however, makes more or less the following claim:

- (C) In certain special cases, a functional kind *F* can be realized even when the token that realizes it is such that its causal relations fail to map precisely onto *F*'s proprietary relations (as specified by the relevant functional theory).

The basic strategy of the objection is to insist that the System Argument exploits one of those 'special cases', and that consequently Sam continues to experience pain at t_2 in spite of the incapacitation, with no harm to functionalism.

The various forms of the objection differ from one another in how the class of special cases is conceived. Special cases might be those in which tokens are of a type that *typically* bear the appropriate causal relations, or *tend to*, in which *ceteris* is not *paribus*, in which conditions are *abnormal*, or *non-ideal*, and so on. With respect to the System Argument, and depending on which form of the objection is adopted, the claim might

be that the incapacitation of Sam's action-planning region at t_2 has no effect on his experience at that time because conditions are abnormal, because the activity in his pain region is of a type that typically causes action-planning region activity, etc. No matter which version of the objection is employed, therefore, the upshot is that functionalism handles perfectly the intuition the argument evokes.

According to the objection, the System Argument misrepresents functionalism by ignoring some aspect or component of functionalist theories that renders them compatible with the intuition of the System Argument. There is some truth to this charge: thus far my account has overlooked certain versions of functionalism, and for two reasons. First, not all brands of functionalism appear to have a component to handle mismatches between a system's interrelated states and processes, and the causal relations specified by the functional theory; and I wanted the System Argument to apply to those which do not.¹⁴ Second, different functionalist theories characterize the added component in diverse ways, as the above paragraph illustrates. For the sake of exposition, therefore, I thought it best to deal with the various construals of the added component separately, after first presenting the basic argument.

In replying to the objection, I shall sketch three ways of spelling out (C)—the claim that a functional kind *F* can be realized even when the causal relations of the realizing token do not match *F*'s proprietary relations. These three ways of interpreting (C), I believe, exhaust the functionalist's options. After offering each construal of (C), I shall argue that it fails to save the functionalist from the System Argument.¹⁵

¹⁴ Certainly there is often no mention of such a component in accounts of functionalism. Moreover, such a component may be incompatible with certain brands of functionalism—namely, those which allow that elements in the world might, accidentally, and for a brief time, become so interrelated as to constitute a system that realizes a functionalist theory. I have in mind examples like Block's (1978) Chinese Nation, Searle's (1990) molecules in walls, and others from the literature. The difficulty with such examples lies in distinguishing systems of interrelated elements that do not realize a given functionalist theory, on the one hand, from those that do, when there is a mismatch with *F*'s proprietary relations *and* conditions are abnormal (*ceteris* is not *paribus*, etc.), on the other. It is just hard to see how such a distinction can be grounded if the relevant elements have entered into their particular causal interrelations *by chance*.

¹⁵ (C), as stated, suggests that special cases are appealed to whenever functionalists claim that *F* can be realized in spite of a mismatch between the causal relations of *F*'s realizing token, and *F*'s proprietary relations as specified by the functional theory. However, it may be thought that this is not so of Lewis's (1972) version of functionalism, and that consequently he has another way of responding to the System Argument. Here is Lewis's recipe for specifying functional theories (in Block, 1980a, p. 212):

Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses . . . Form the conjunction of these platitudes; or better, form a cluster of them—a disjunction of all conjunctions of *most* of them. (That way, it will not matter if a few are wrong.)

On the version of functionalism that emerges, *F* can be realized in spite of a mismatch between the causal relations of *F*'s realizing token, and *F*'s proprietary relations as

Consider a version of functionalism on which a token must bear at least *some* of a functional kind F's proprietary relations for F to be realized—and even under abnormal conditions, when *ceteris* is not *paribus*, etc. (Where typical causes are at issue, the proposal would be that for some of F's proprietary relations, it is not enough to realize F that a token be of a type that typically bears those relations; it must actually bear them.) There are two ways of understanding this, and they correspond to two of the three construals of (C) I shall discuss. According to the first, only certain of F's proprietary relations are such that they can be lacking and F still be realized; other of F's proprietary relations are such that a token must have them to realize F. Take, for example, a functional theory on which pain is, among other things, a state that causes the construction of a plan to fetch an aspirin (given certain input conditions, etc.). It might be held that (under those same conditions) a pain can be realized even if it is not disposed to cause an action-plan construction—when circumstances are abnormal, say—but that a pain cannot be realized *under any circumstances*, abnormal or otherwise, if certain of pain's other proprietary relations are lacking—say, a relation to a belief that one has pain. On this view, abnormal conditions, tendencies to cause, *ceteris paribus* conditions, and the like, apply only to some, and not all, of a kind's proprietary relations.

It should be clear that any functionalism along these lines cannot defend against the System Argument. For the argument simply can be run on any of the relations that a token must bear, according to the theory, to realize F. If being causally related to an action-plan construction is unnecessary for pain to be realized (under abnormal conditions, say) then one need only substitute a relation that *is* required (even under abnormal conditions) and the argument goes through without a hitch. Since, as we often have had occasion to note, the System Argument can be run with any conscious state or process whatever, as well as with any effect, making

specified by the conjunction of the platitudes. No appeal to special cases (e.g. abnormal conditions) is required, but only to the fact that most of F's proprietary relations are realized. Now it may seem that this suffices to answer the System Argument. It does not, however; for we are focusing on the wrong sort of mismatches. We must consider mismatches that exist between a system's interrelated states, *and the entire disjunction of conjunctions*—which, after all, is what properly expresses the functional theory. Such mismatches can be had by starting with a system that satisfies the relations specified in *only one* of the disjuncts of the theory, and then eliminating one or more of those relations. The System Argument works perfectly well when run on such a system. If one then wants to claim that such a system still realizes F in spite of that mismatch, it seems one must appeal to the sort of special cases mentioned in (C). Lewis, for example, would likely invoke his notion of typical causes (Lewis, 1966; 1980). The rest of this section can be seen as answering that kind of move. (To apply what is said in this section to Lewis's (1972) version of functionalism, take the systems that are discussed to realize, at most, the relations specified in one of the theory's disjuncts; and construe the three interpretations of (C) as pertaining only to the relations specified within that disjunct.)

the substitution is unproblematic. The first way of elaborating (C), therefore, is of no help to the functionalist.

A second way of interpreting the claim that a token must bear at least some of F's proprietary relations to realize F is this. Instead of maintaining that certain of F's proprietary relations are such that a token must have them to count as F, one might hold a 'cluster' view of sorts, according to which a token must bear a sufficient number (or summed weighting, etc.) of F's proprietary relations to realize F, though no specific relation is such that a token must bear it. This option, however, fares no better than the first. To see why, imagine the System Argument altered as follows. Before Sam's action-planning region is incapacitated at t_2 , suppose that his brain is already damaged, and to such an extent that the causal relations had by the activity in his pain region map onto a *just sufficient number* of pain's proprietary relations for pain-region activity to count as pain, according to some cluster view. It follows that if Sam's action-planning region is incapacitated at t_2 , when his pain region is still activated, his pain will cease, since no longer will a sufficient number of pain's proprietary relations be intact. However, so long as Sam's action-planning region is sitting idle at t_2 (because the signal from his pain region has not yet arrived), the idea that incapacitating it at t_2 should affect Sam's experience *at that time* is as incredible as before. The 'cluster' view under consideration, therefore, fails to avert the argument.

If a token must bear some of F's proprietary relations to realize F, then either specific relations must be had, or any will do so long as (roughly) the token bears enough of them. Since both options have been found wanting, it would appear that the only alternative left to the functionalist is an interpretation of (C) that implies the following: under the right conditions (abnormal, non-ideal, etc.), F can be realized even if the realizing token bears *none* of F's proprietary relations. This means that in certain circumstances tokens can realize functional kinds *entirely outside any containing system*. In Sam's case, it means that his pain region could be removed from his brain, activated, and a token experience could be realized. All that would be required is that the conditions be abnormal, that Sam's activated pain-region be of a type that typically bears the right relations, and so on.

What is to be made of this? The first thing to notice is that versions of functionalism that take (C) in this manner have a response to the System Argument—unlike versions based on the first two construals, which are unable to prevent the System Argument from getting a foothold. If abnormal conditions, *ceteris paribus* clauses, and the like can be applied when any combination of any of F's proprietary relations are lacking, it will always be possible to accommodate the intuition of the System Argument, so long as it can be argued that abnormal conditions, etc. have been exploited. Accordingly, I believe that the best response to this third construal of (C) is one that shows that the variety of functionalism that is engendered is unacceptable.

Fortunately, this does not seem difficult. Aside from the fact that the view looks false on its face—since it seems obvious that as one destroys more and more of a system (e.g. a brain), tokens that realize conscious states will cease to realize anything mental well before they are the sole remaining parts of the system—there are two reasons for rejecting any functionalism along these lines. First, if physical tokens like ‘activated pain regions’ can realize psychological states like pain in complete isolation from brains or other systems, it is hard to see why one would even consider functionalism to be correct, when a neurophysiological type–type identity theory (involving a long disjunction perhaps), for example, would be more natural.¹⁶ Espousing a functionalism on which mental states can be destroyed only if the tokens that realize them are, goes a long way toward undermining any motivation for thinking functionalism true in the first place.

The second reason is as follows. On the version of functionalism being considered, isolated tokens can in certain special cases realize conscious kinds. Sam’s pain region, removed from his brain and activated, for example, might realize a pain if it is of a sort that typically causes action-planning region activity, if *ceteris is not paribus*, etc. However, it seems entirely possible that there could be a molecule-for-molecule duplicate of Sam’s activated pain region for which being isolated from a brain is a *normal* condition. Such a duplicate, for instance, might never have been part of a brain, it having evolved for some unrelated purpose, if at all.¹⁷ Now because the duplicate lacks pain’s proprietary relations, and because conditions are *normal* (*ceteris is paribus*, etc.), the functionalist must deny that the duplicate realizes pain. However, that amounts to denying that types of conscious experiences supervene on the physical make-ups of whatever have them. And that, I assume, is unacceptable. While one might wish to argue that supervenience fails for the semantic contents of physical states and processes, that conclusion would appear to be unavailable for consciousness.¹⁸

¹⁶ This first point does not apply to Lewis’s (1966, 1972, 1980) brand of commonsense functionalism, which enables him concurrently to hold a type–type identity theory. However, the second point (in the next paragraph) does apply to Lewis, since for him, so far as I can tell, *all* causal relations among inputs, outputs, states and processes get specified in the theory in terms of typical causation.

¹⁷ If one would prefer a different example, imagine a wire in a robot’s brain that realizes pain when it is in some physical state *S*. Though conditions might be abnormal for the wire if it were to be removed from the robot, conditions could surely be normal for a ‘twin wire’ in state *S* that is wrapped around a spool in a hardware store.

¹⁸ There is a related argument that can be wielded against the class of objections as a whole. Consider any story on which Sam can realize pain in spite of the fact that activity in his pain region does not bear all of pain’s proprietary relations. (Or on Lewis’s (1972) version of functionalism, in spite of the fact that activity in his pain region does not bear all the relations specified in any of the disjuncts. See n. 15.) Any such story will have it that Sam is in pain partly because conditions are

This third way of construing (C), then, results in an implausible version of functionalism. Since the first two ways give rise to versions that are incapable of blocking the System Argument, the functionalist's options would now appear to be exhausted. I conclude that the objection considered in this section poses no threat to the System Argument.

6. Conclusion

I believe the conclusion to be drawn at this point is this: either functionalist theories of consciousness are false, or something is wrong with the intuition underlying the System Argument. As regards the intuition, one might simply deny having it; or alternatively one might hold that although one has it, the intuition nevertheless is an illusion. Either way the upshot is that tampering with idle system parts can affect a system's experience, as functionalism dictates. (Notice that one cannot both have it this way, and appeal to the objection discussed in the previous section.) Now I have little to say, on top of what I have said already, to one who claims not to have the intuition. And anyone who maintains that the intuition is an illusion owes a non-question-begging account of why it is an illusion, and how the illusion arises.¹⁹ Offhand I know of no plausible account that can be offered. Not finding any reason to treat the intuition as suspect, therefore, I conclude that functionalism with respect to consciousness is false.

Department of Philosophy
Tel Aviv University
Tel Aviv 69978
Israel

abnormal, etc. But it would seem that there could always be a physical duplicate of Sam, Sam*, for whom conditions are *normal* (*ceteris paribus*, activity in his pain-region-analog does not typically bear pain's proprietary relations, etc.). Sam* would not experience pain when Sam does, even though they are physical duplicates. Types of experiences, therefore, would fail to supervene on the physical make-ups of individuals that have them.

- ¹⁹ One might claim, for example, that the intuition of the System Argument ultimately rests upon intuitions about qualia; but qualia do not exist; hence the illusion. Both of these points must be argued for, however. Arguing for the first, I suspect, would be no simple matter. Why, for example, should one suppose that the intuition driving the System Argument involves intuitions about *qualia*, rather than simply about *consciousness*? 'Qualia', after all, is a term of art, and it is not employed in any of the arguments in this paper (outside of this note). Arguing for the second point (that qualia do not exist), on the other hand, would require that one make quite specific claims about what qualia are supposed to be—which would in turn make arguing for the first point that much more difficult.

References

- Antony, M.V. 1990: *Consciousness, Content, and Cognitive Architecture*. MIT Ph.D. thesis.
- Antony, M.V. 1992: The Where and When of What? Commentary on Dennett and Kinsbourne (1992). *Behavioral and Brain Sciences*, 15, 201–2.
- Block, N. 1978: Troubles with Functionalism. In C. Wade Savage (ed.), *Perception and Cognition, Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science, Volume 9*. Minneapolis: University of Minnesota Press, 261–325. Reprinted in Block (1980a).
- Block, N. (ed.) 1980a: *Readings in the Philosophy of Psychology, Volume 1*. Cambridge, MA.: Harvard University Press.
- Block, N. 1980b: Introduction to Functionalism. In Block (1980a).
- Block, N. 1993: Review of D.C. Dennett, *Consciousness Explained*. *Journal of Philosophy*, 90, 181–93.
- Dennett, D.C. 1991: *Consciousness Explained*. Boston: Little, Brown and Co.
- Dennett, D.C. and Kinsbourne, M. 1992: Time and the Observer: The Where and When of Consciousness in the Brain. *Behavioral and Brain Sciences*, 15, 183–248.
- Lewis, D. 1966: An Argument for the Identity Theory. *Journal of Philosophy*, 63, 17–25. Reprinted in Lewis (1983).
- Lewis, D. 1972: Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy*, 50, 249–58. Reprinted in Block (1980a).
- Lewis, D. 1980: Mad Pain and Martian Pain. In Block (1980a). Reprinted in Lewis (1983).
- Lewis, D. 1983: *Philosophical Papers, Volume 1*. New York: Oxford University Press.
- Loar, B. 1981: *Mind and Meaning*. Cambridge University Press.
- Maudlin, T. 1989: Computation and Consciousness. *Journal of Philosophy*, 86, 407–32.
- Rey, G. 1993: Sensational Sentences. In M. Davies and G.W. Humphreys (eds), *Consciousness*. Cambridge, MA.: Basil Blackwell, 240–57.
- Searle, J. 1990: Is the Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association*, 64, 21–37.
- Shoemaker, S. 1981a: Some Varieties of Functionalism. *Philosophical Topics*, 12, 83–118. Reprinted in Shoemaker (1984).
- Shoemaker, S. 1981b: The Inverted Spectrum. *The Journal of Philosophy*, 74, 357–81. Reprinted in Shoemaker (1984).
- Shoemaker, S. 1984: *Identity, Cause, and Mind*. Cambridge University Press.