Consciousness, Value and Functionalism

William Seager
Philosophy Division of Humanities
University of Toronto at Scarborough
Scarborough, Ontario, M1C 1A4
CANADA

seager@utsc.utoronto.ca

PSYCHE, 7(20), November 2001
http://psyche.cs.monash.edu.au/v7/psyche-7-20-seager.html

KEYWORDS: consciousness, value, functionalism, token-identity theory, qualia

COMMENTARY ON: Charles Siewert. (1999). *The Significance of Consciousness.* Princeton University Press. 392 pp. ISBN: 0691027242. Price: $US 42.50 hbk.

ABSTRACT: Charles Siewert presents a series of thought experiment based arguments against a wide range of current theories of phenomenal consciousness which I believe achieves a considerable measure of success. One topic which I think gets insufficient attention is the discussion of functionalism and I address this here. Before that I consider the intriguing issue, which is seldom considered but figures prominently at the close of Siewert's book, of the *value* of consciousness. In particular, I broach the question of whether the value of consciousness has any impact on our theoretical understanding of consciousness.

1. The Value of Consciousness

I find myself in the awkward position, for a philosopher, of agreeing in large measure with the target of my article. I think that current physicalist theories of consciousness *are* open to the objection that they "neglect" the core feature of consciousness, which is its qualitative or experiential aspect. I'm not sure, however, that "neglect" is the right term for this; most proponents of these accounts of consciousness know only too well what they are supposed to explain and they do try to explain it. But it's not clear that they succeed, and in large measure Charles Siewert's book, *The Significance of Consciousness,* is a powerful series of arguments, concerning a variety of theories of consciousness, which aims to reveal their failure.

So instead of launching into the traditional unremitting criticism of a fellow philosopher, I want to extend two of Siewert's themes: the intrinsic value of conscious experience, which opens some new territory in consciousness studies, and Siewert's attack, which I see as underdeveloped, on functionalism.

The last chapter of *The Significance of Consciousness* is devoted to what might be regarded as the *real* significance of consciousness: its value. The core thought here is that conscious experience is *intrinsically* valuable, where this means that there are some conscious experiences which are worth having for themselves and not because of anything else they might produce or lead to. The sense of value which can be ascribed to conscious experiences in and of themselves, and no matter their consequences, was beautifully expressed by Dostoevsky, where, near the beginning of chapter 5 of Part 2 of *The Idiot,* he has Prince Myshkin reflect that:

> He remembered that during his epileptic fits, or rather immediately preceding them, he had always experienced a moment or two when his whole heart, and mind, and body seemed to wake up to vigour and light; when he became filled with joy and hope, and all his anxieties seemed to be swept away for ever; these moments were but presentiments, as it were, of the one final second (it was never more than a second) in which the fit came upon him. That second, of course, was inexpressible.

Myshkin goes on to ponder the material underpinnings of this state of consciousness:

> What matter though it be only disease, an abnormal tension of the brain, if when I recall and analyze the moment, it seems to have been one of harmony and beauty in the highest degree - an instant of deepest sensation, overflowing with unbounded joy and rapture, ecstatic devotion, and completest life?

And finally concludes:

> Since, in the last conscious moment preceding the attack, he could say to himself, with full understanding of his words: "I would give my whole life for this one instant," then doubtless to him it really was worth a lifetime.

I think everyone can recognise the kind and degree of value which certain states of consciousness possess in and of themselves which this passage so strikingly evokes.

The axiological significance of consciousness has not been looked at very closely by philosophers of mind and only indirectly in value theory so Siewert's raising of the issue is important and original. Recall that the founding utilitarian, Jeremy Bentham set *pleasure* - a state of consciousness - as the fundamental good; John Stuart Mill was less restrictive, opting for *happiness,* but this is no less a "state of mind" which is certainly to be interpreted as a state of consciousness. G. E. Moore also regarded states of consciousness as the possessors of intrinsic value. More recently, the champions of "deep

ecology" have tried to extend our notion of intrinsic value to include nature, eco-systems, biological diversity and other utterly non-conscious entities, but so far as I know the deep ecologists have never tried to deny the intrinsic value of consciousness (and I think few find their attempted extension of value unproblematic).

Siewert argues that states of consciousness possess intrinsic value and his arguments seem to me utterly convincing. I think Siewert's usual practice of carefully detailing pure philosophical thought experiments is especially appropriate for this question (and I should mention that Siewert's use of careful and detailed thought experiments is a welcome reminder, at a time when some have questioned the legitimacy of their use, of how powerful a method of philosophical argumentation thought experiments can provide). This is because questions of value assignment are independent of the kind of possibility which the thought experiment depends upon or exploits. We can assign, if only hypothetically, value to things that are physically impossible or even metaphysically impossible (perhaps even to things that are logically impossible - I might wish that arithmetic had not "turned out to be" incomplete for example).

Consider the thought experiment invoked to show that phenomenal consciousness has intrinsic value (pp. 319 ff.). The experiment involves imagining a choice between a life with and a life without consciousness (Siewert calls the second option "zombification"). It is a large and contentious issue in philosophy of mind whether zombification is possible and, if so, what sort of possibility it possesses. But whether or not it is physically, or metaphysically, possible for all one's behavioral and perceptual-discriminative capacities to be preserved despite the loss of consciousness, we can still understand the proposal well enough to determine the relative value of each (at least, to me this is trivially easy).

Imagine the devil gives you the choice: you can become the richest and most successful person on the planet, but at the cost of a total loss of consciousness. You will be a zombie, though undetectably such to the rest of the world, and a very well off zombie at that. It is easy to see that, all other things equal, this offer is no bargain; it is tantamount to death. (Now, this does not mean that no one would choose it. Many people have concerns for which they are willing to die. If by becoming so wealthy I could make serious improvements in the state of the world or save the lives of my family - and my zombie version, we can stipulate, will not lose my so to say "ethical dispositions" - then perhaps I would be willing to accept zombification. But obviously I would rather do all these good things consciously.)

Perhaps a more interesting claim, or at least a claim that is much more difficult to demonstrate, is that the *only* things which possess intrinsic value are states of consciousness. I believe that this is true but don't know how to prove it. One might argue from the premise that there is no value without *valuings* and that these latter states have to be conscious. But I don't know how to show that this premise is true either.

An obvious thought experimental approach is to imagine universes in which consciousness does not or even *cannot* exist. Consider a range of such universes differing

in a host of various properties. That is, imagine worlds differing in their basic laws and constants of nature and hence differing in their emergent properties, such as whether stars exist or not, whether the universe collapses back into a singularity within a microsecond or not, etc. The experiment requires that in all the worlds under consideration, it is impossible for consciousness to exist or to emerge in them. With this body of universes in mind, let us try to rank them in terms of intrinsic value. Which universe would be best? They all seem equally and totally empty of value in themselves. We might contend that those universes with the maximal diversity springing from the minimum number of basic principles are the *best* (this was Leibniz's criterion but of course his worlds were shot full of consciousness). But why should anyone think that diversity has intrinsic value?

The general problem here is that although it is easy to see why states of consciousness are intrinsically valuable since they have a subjective component that, so to speak, directly reveals their value (see Prince Myshkin's ruminations above), it is impossible to see *why* any non-conscious state should have intrinsic value. One can, of course, *assert* that diversity, say, is intrinsically valuable, but unless diversity is tied to some conscious being's *appreciation* of diversity, which it almost invariably is, or some other valued *effect* of diversity the assertion seems completely empty and devoid of support.

Turning away from the grandiose thought experiments involving the creation of whole universes, we can perform microworld thought experiments as well. Consider the popular computer game, *The Sims*. I'm sure you are familiar with it. I hope you agree that, although enjoyable and perhaps worthwhile for those playing it, the sims themselves and their worlds (which are just little neighbourhoods) do not matter at all. But imagine that, somehow, we discovered that the little functional architecture which underlies the onscreen behaviour of each sim was sufficient to generate a small measure of consciousness. Imagine that the sims really do feel hungry sometimes and uncomfortably in need of the bathroom at other times (see Stanislaw Lem 1980 for an intricate and amusing version of this thought experiment). It is obvious that once I know that the sims are conscious creatures, there is a real moral issue about whether or not I'm not allowed to turn off the game, because once I know the sims are conscious I know that their states *matter*. And it seems to me that nothing else I could learn about the sims and their little worlds would, or even could, make their states matter in themselves.

Whether or not it is true that *only* conscious experiences have intrinsic value, it is clear that some things which have value do *not* have intrinsic value. For example, the taking of aspirin is good if you have a headache and it is good just because of its effects on your consciousness. Everything which has value which is not intrinsic value ultimately has value because of its relation to something which does have intrinsic value. This sort of value is called instrumental value (the notion is recursive, so some things have instrumental value because they help produce other things with instrumental value, but eventually value is grounded in what is intrinsically valuable). To return to the sims (that is, the actual ones, not my imaginary conscious ones), I could learn that eliminating a particular sim might set off a bomb somewhere; this would give value to that sim's continued existence. This shows that *anything* can have instrumental value simply depending on the circumstances in which it exists.

The fact that states of consciousness have intrinsic value, when combined with the now generally accepted doctrine of the multiple realizability of mental states, leads to a rather curious argument which I would like to present briefly here. I think this at least shows that the axiological significance of consciousness which Siewert emphasizes has potentially interesting ramifications in our traditional, non-axiological, philosophy of mind.

In brief, multiple realizability implies that brain states (assuming that brain states are the states which realize our states of consciousness) have merely instrumental value. Consider, for example, the neurosensory cells - the hair cells - within the cochlea which translate motion (caused by currents in the fluid of the inner ear which are in turn caused by rapid variation in air pressure, or sound waves) into electrical signals which feed into the auditory nerve. Although the hair cells are in all likelihood too peripheral to be realizers of conscious states, I pick this as an example since we know that the hair cells' function can be carried out by a purely electrical device known as a cochlear implant. These devices can restore hearing to those who have damaged or destroyed hair cells. I take it as now agreed that the conscious experience of (at least some kinds of) sound has intrinsic value (music is only one example). But it does not matter whether the valuable experience is generated by hair cells or an implant. The fact that this does not matter means that the job of translating sound waves into neural electrical signals has only instrumental value. That is, all other things being equal, it doesn't matter to you whether your experiences of sound are brought about by hair cells or a cochlear implant. (Of course, all things are not equal: implants - certainly not current ones - probably cannot match our neural machinery in the range and subtlety of experiences produced, there are medical issues about implantation, etc.)

In general, the multiple realizability of states of consciousness means that, all other things equal, it doesn't matter how these states are realized. Suppose the proverbial evil alien neuroscientists abducted me and replaced half my neurons with silicon isomorphs last night. By definition these replacement neurons are functionally identical to the neurons they replaced. For the sake of the argument we can assume that it is those functional properties of the neurons which the replacements duplicate - their input/output "program" say - which matters to the implementation of conscious states (as indeed many believe and may well be the case). Multiple realizability entails that, given these assumptions, the states of consciousness which occur post-abduction are, in general, qualitatively no different, as states of consciousness, from those which occurred pre-abduction. So these states retain their intrinsic value as states of consciousness. It just does not matter how they are being implemented, and, all other things being equal, I wouldn't care that my brain had been altered. But I most certainly do care that states of consciousness are being implemented in me (that is what it means to say that they have intrinsic value). This can be summed up in the claim that implementing states have instrumental value; their value lies in the fact that they realize states (of consciousness) which have intrinsic value.

Now we can use Leibniz's law to produce a very simple argument to show that brain states are *not* identical to states of consciousness:

1. States of consciousness have intrinsic value.
2. Brain states only have instrumental value, hence do not have intrinsic value.
3. Therefore, brain states are not states of consciousness.

I think this argument can be extended to explicitly refute even the token identity of states of consciousness and their realizing brain states. Consider again the silicon isomorph neural replacement thought experiment. It seems to me that I would not be destroyed by the operation which replaces my biological neurons with silicon counterparts; I would retain my identity despite the changes in my brain. More controversially perhaps, I think that whether or not the aliens perform the operation, I will have numerically the same thoughts the next day (here I assume determinism for simplicity but it is not really crucial to the argument). That is, I deny that the realizing states are essential to the identity of the states they realize. Suppose that, as I trust is indeed the case, there are no such aliens and today my states of consciousness are actually subserved by my good old biological neurons. At 9 am this morning I had the conscious thought: "spring has finally arrived". Consider, now, the counterfactual claim that *if* the aliens had replaced my neurons with silicon isomorphs last night, I would have had the *same* thought at 9 am this morning. Given determinism, it is obvious that I would have had at least *a* conscious thought that spring has finally arrived at 9 this morning. But I think we can conclude more than this. I think that this would have been the very same thought and not just another instance of the "spring has arrived" kind of thought. That is, I think it is correct to say things like: *this* very thought might have been realized with silicon isomorphs.

Token identity is refuted because obviously it is impossible for the activation of silicon isomorphs to be (or to have been) the activation of biological neurons.

Note that the argument does not generalize indiscriminately. We might say that my desk is "realized" by a set of wooden parts standing in certain relations to each other. But we *cannot* say that *this* desk might have been made out of plastic although we may allow that *a* desk just like it, to all appearance, could have been made of plastic and we can allow that my desk might have been a plastic desk (this last is a different claim altogether in fact). The lesson we should draw is that the realization relation is not part of mereology (my states of consciousness do not have neural activity as parts).

I will not pursue the argument any further here. In a way, the main point is just to illustrate that the idea that states of consciousness have intrinsic value can do some work outside of areas concerned with purely axiological issues.

Instead I want to turn to Siewert's attack on functionalist accounts of consciousness, which I find to be somewhat underdeveloped.

2. Functionalism

Recall that Siewert makes a distinction between "manifest feature functionalism" and "hidden feature functionalism". Functionalism is characterised, following Ned Block, as defining the notion of a mental state so that "each type of mental state is a state consisting of a disposition to act in certain ways and to have certain mental states, given certain sensory inputs, and certain mental states". Manifest feature functionalism adds that "the [relevant] types of mental states, acts and sensory inputs are types we can claim with warrant we instance without internal examination" (p. 141). In contrast, hidden feature functionalism asserts that there are internal feature unavailable to us without "internal examination" which make "the difference between having and lacking conscious visual experience" (pp. 145-6).

With regard to manifest features, Siewert claims of Connie (visually conscious but visually impaired) and Belinda (blindsighted but capable of the same visual discriminations as Connie) that "the only differences in manifest features they had or were disposed to have, were such as cannot rightly and non-trivially identify ... with the difference between having and lacking ... visual consciousness" (p. 141). But of course one of the primary differences just is that Belinda is given to saying things like "but I don't *see* anything" even as she successfully discriminates visual stimuli just as well as Connie. Such denials of visual consciousness will obviously be a manifest feature of the functional organization - broadly speaking - of her visual system which distinguishes her from Connie. We cannot allow that this possible difference between Belinda and Connie is irrelevant to whether or not they possess visual consciousness. True, it is not a difference which could *constitute* the distinction between being conscious visually and not being so conscious, but it is clearly the kind of behavioural evidence which we actually do use to tell whether or not a putative blindsighter does have (only) blindsight. Siewert seems to deny this on p. 141 where he asserts "the point remains that we cannot make a conscious visual experience purely out of what Belinda has, plus a propensity, which only Connie would have, to think that one has such an experience". I don't think this is obvious. Suppose that it is impossible to think that one is enjoying visual consciousness if one is not visually conscious (a version of infallibility about consciousness: if one thinks one is (visually) conscious then one *is* (visually) conscious). It is not obvious that this principle is false, although its converse likely is since we intuitively suspect that animals, for example, can be visually conscious without having any introspective abilities. It is hard to think of any counterexamples to the principle which aren't wildly fictitious (I am taking it that in dreams and imagining one can be and normally is "visually conscious") and hence rather tendentious.

If the principle is correct, then, assuming that Belinda is sincere in her assertions, an assumption we are surely permitted to make in this case, then adding Connie's disposition to Belinda would seem *strictly* to entail that she has visual consciousness. What more could you want?

Of course, it is not by "magic" that adding the proper dispositions generates visual consciousness. Without a substantial account of *why* the addition is sufficient to produce experience, we are in the situation of the possessor of the "ultra-shoes", which Siewert derides on pp. 131ff. Ultra-shoes are shoes which are exactly like ordinary shoes save for

their ability to induce in their owners that these shoes are indeed ultra-shoes. But in Belinda's case we know that adding the disposition will require more or less extensive and general alteration of a host of neural pathways and processes. Presumably, it is impossible to make these alterations without engendering consciousness. The status of the relation between the addition of the relevant dispositions and the creation of consciousness is somewhat analogous to the relation between the general law that perpetual motion machines are impossible and the particular details of any one machine. We know that no machine can be a perpetual motion machine, but there is no general recipe for specifying where a proposed perpetual motion machine will conflict with physical law. Similarly, although we have no idea what details of brain operation will both have to be changed to grant Belinda the requisite dispositions *and* will underwrite visual consciousness, but the above principle of incorrigibility guarantees that these details must exist. (This point also interacts with the issue of hidden-feature functionalism, for which see below.)

One might also object that the disposition to utter something is not equivalent to believing what is uttered. True enough, but the disposition under consideration is much more complex than merely a propensity to utter a few sentences; it involves many behaviours and many kinds of utterances. If Connie believes she is visually conscious (as she most certainly does) then adding her dispositions to Belinda will surely result in Belinda believing that *she* is visually conscious, and if believing that one is visually conscious entails that one is conscious then the functionalist succeeds.

We are forced to consider more closely the claim that it is possible to believe that one is visually conscious while being entirely visually unconscious. If this is possible then the links between visual consciousness and beliefs about visual consciousness have been severed. If these links are severed (or are severable) then the absence or presence of such links will not count as part of the "manifest functional role" of states of visual consciousness. And this means that no functionalist ought to think that adding Connie's disposition to think that she is visually conscious to Belinda would add visual consciousness if it was lacking before.

Arguments like Siewert's (and many others) have the same feature: duplicate the FFC (full functional characterization) of a conscious state, a, in a distinct state, b, while denying that the consciousness of b (or associated with b, or involved with b - I do not want to imply that a state is conscious only if someone is conscious *of* it) is identical to that of a. Thus the argument goes that although we can add to Belinda enough to make her current state fully functionally identical to Connie's, we shall not thereby make Belinda's state a state of visual consciousness.

Note also that once we sever the link between visual consciousness and beliefs about visual consciousness we open the possibility that so-called blindsighters retain their visual consciousness but lose the normal links between it and behaviour and belief. That is, the assumption that it is possible that there is the kind of spontaneous blindsight which Siewert hypothesizes assumes that this sort of blindsighter would be lacking in visual consciousness. This assumption gains its plausibility from "ordinary" (actually existent)

blindsight where we have huge manifest functional differences between sighted and blindsighted, of precisely the kind that indicate lack of visual consciousness. Why shouldn't we believe that the more peculiar forms of blindsight invented by Siewert are (or, since they are entirely imaginary, could be construed as) just a weird kind of visual consciousness coupled with a range of behavioural and memory deficits? It comes dangerously close to begging the question against functionalism to make this assumption and never defend it. How do we know that by adding to the blindsight thought experiments we are not surreptiously introducing creatures that are visually conscious, most especially in the purely imaginary case Siewert labels spontaneous blindsight. If we can, by mere philosophical fiat, make it so that such blindsighted creatures are possible then of course functionalism must be false but then the weight of the argument rests on this possibility claim itself.

Perhaps this is a bedrock clash of intuitions. Siewert thinks that blindsight could remain as we make Belinda's cognitive system a FFC of Connie's; a functionalist thinks that the convergence of the FFC just is the birth of visual consciousness in Belinda. This would not be the denial of phenomenal consciousness that we find in some of the possible responses to which Siewert considers against his own thought experiments. Blindsight is possible, but not a kind of blindsight that refutes functionalism. Is there any independent *argument* that a functionalism refuting blindsight is so much as possible? Or do we just have the intuition that it *must* be possible to extend ordinary blindsight into the realm of functionalism refuting thought experiments?

Thus it may be that Siewert's thought experiments stand with all the other anti-functionalist thought experiments (such as Searle's (1980) "Chinese room", Block's (1978) "Chinese nation" etc.). Siewert's thought experiments do have the advantage of being, on the face of things, considerably less "weird" than the traditional ones. But they end up weird enough. What Siewert calls spontaneous blindsight has never been observed and the reason seems clear enough - blindsighters aren't visually conscious so there is nothing to trigger a spontaneous reaction to the visual information which is somehow entering their cognitive systems.

So although I am intuitively sympathetic to Siewert's conclusions about functionalism, the arguments on each side here seem to lead to a standoff.

However, if there are belief states which are non-conscious then it seems that no functionalist would actually endorse *manifest* functional role functionalism since a non-conscious system could believe, with warrant too I guess even though it was wrong, that it was conscious. The only way to find out, even for the system itself, that such a system is not a conscious as well as a thinking being would be via an internal examination. Thus we really must turn to "hidden feature" functionalism to find a plausible doctrine to attack. It is the nature of the inner states that cause the behaviour by which we ascribe consciousness and thought that must make the difference, and the functionalist characteristically asserts that what matters in the causal organization of this set of states and their propensity to cause certain kinds of behaviour, and certain other such states, and be caused by certain sorts of sensory stimulation, and certain other such states.

One point that might engender confusion is this. Functionalism can be divided into what I will label analytic and scientific forms. Although the former has some affinities with manifest feature functionalism it is nonetheless quite distinct. It asserts that our concepts of mental states are concepts of states which play a certain causal role in the production of other mental states and behaviour, and in being produced by other mental states and sensory stimulation. The states themselves are not manifest states, and the nature of the roles precludes certain sorts of inner states from constituting mental states. The scientific form of functionalism differs in that it just doesn't care whether or not our present concepts of mental states are functional concepts; it enjoins us to reform our understanding so that we begin to conceive of mental states in functionalist terms (the payoff being scientific understanding, various sorts of cognitive "technology" and a proper appreciation of what the mind could be from within the scientific picture of the world).

Obviously scientific functionalism has resources available to distinguish Belinda from Connie which analytic functionalism lacks insofar as the former can assign functional roles to mental states that are entirely internal and so to speak invisible to the system. For example, one hypothesis that has been entertained to account for blindsight is that the condition results from the destruction of a "high level" (perhaps serving more abstract, conceptual cognitive functions) information pathway from the visual cortex to various other brain regions combined with the preservation of some of the several "low level" pathways (perhaps serving more basic, action oriented functions). Quite a few information/consciousness dissociations are now known; see Rossetti (1998) and a variety of different neural pathways have been proposed to account for them. Perhaps the crucial functional linkages that underpin visual consciousness reside in a particular neural pathway, and their loss naturally destroys consciousness while the preserved existence of other neural pathways explains the persistence of some visual *information* which the brain, and the subject, can still access under special circumstances.

One must regard an hypothesis such as this as fitting into the functionalist picture insofar as one allows that it is the organization of certain parts of the brain which, by hypothesis here, underlies visual consciousness. An extreme test of this functionalist reading is the intuitive response to the question whether visual consciousness could be duplicated in non-neurological material that preserves the appropriate organization. For example, we might imagine, as in any number of philosophical thought experiments, and such as some advanced above, in which real neurons in the appropriate regions of the brain are replaced with artificial silicon neurons which preserve the brain's connectivity. One is pretty strongly inclined to think that this would not affect consciousness in the subject, and thus if one is any kind of physicalist about the mental (and about consciousness in particular) then one appears to be a kind of (hidden feature) functionalist about it as well. This seems a strong reply to the first line of argument against hidden feature functionalism which Siewert deploys on p. 147.

Siewert urges two other arguments against hidden feature functionalism. One is just an appeal to the abstract metaphysical possibility of zombies which despite being physically identical to conscious beings (and hence trivially functionally equivalent as well) are

totally unconscious, along the lines of argument presented by David Chalmers (1996). Zombies are a problem, but they are a problem for physicalists in general, not just functionalists. If zombies are possible then none of the properties of phenomenal consciousness can be physical properties, since the zombie is *ex hypothesi* physically identical to, or shares all physical properties with, a being with such properties, yet the zombie *lacks* consciousness. Therefore every form - even a very weak one - of physicalism is false. That seems a pretty heavy hammer to pound functionalism with (in any case, the determined functionalist *could* respond by saying that the possibility of zombies only shows that brute matter is unable to instantiate the requisite functional roles, though I concede there is something bizarre about such a reply, which seems to involve denying the Church-Turing thesis).

Finally, Siewert asks us how we could ever identify the hidden functional feature which ought to be identified with visual consciousness? Such a question is either easy, in principle, to answer or impossible. It's easy if what we want is the functional structure which is associated with consciousness in the kinds of conscious brains we can observe in the actual world. It is flat out impossible if we want a metaphysical guarantee that "functionally equivalent zombies" cannot exist. I don't think functionalism is in the business of giving such guarantees anyway. Presumably it is a kind of *hypothesis* that mental states are functional states, which can be supported by the kind of evidence we obtain in neuroscience and computer science. This hypothesis is to be gauged as more or less successful insofar as it is explanatory useful, fruitful, elegant and can be integrated into the rest of our scientific hypotheses.

Unfortunately, we have hardly even begun to exploit any of the consequences of this hypothesis and so have very little direct evidence either to support or refute it. We have done some replacement of neural with artificial devices, but only at the extreme periphery of the sensory system. For example, it seems pretty clear (perhaps absolutely certain) that cochlear implants do not leave their recipients partial zombies, that is, able, after the implantation, to respond behaviourally to micro-fluctuations in local atmospheric pressure but still utterly lacking consciousness of them. Consider this lovely passage from a 12 year old implant recipient:

I hope you are all familiar with how the implant works. It's pretty amazing! My brain has learned to interpret the messages sent to it. I can tell you, it doesn't hurt! People often ask me if things sound different. But I can't answer that, because I don't know how things sound to everyone else. They also ask if I hear things right away or if sound is delayed. It's not like a foreign movie that has been dubbed, where the actor's lips are saying one thing, and the voice is saying another. I experience the sound and lip movements at the same time.

When my implant is off, I can't hear anything. Sometimes I like to watch TV without the sound. I imagine the character voices and make up the sounds in my head. I also love captioning, and it helps with TV and movies a lot.

I don't wear the implant when I sleep, so in the morning when I wake up it's a shock to put the magnet on. Whoa! It's like all these sounds come in. It's a blur of sound at first, and it takes me a second or two. Then my brain figures out what the different sounds are: the radio, music, Dad cooking breakfast, things dropping, the traffic outdoors, my parents asking if I remembered this or that, or all of us going over the plans for the day (from http://www.allhear.com)

Of course, no one would be surprised at this since the implant is so peripheral and we all expect that consciousness arises from much more internal processing. Still, it represents the *kind* of thing which will provide evidence in favour of some kind of functionalist account of consciousness; someday we will have lots of "central implants" of various kinds.

To sum up, the arguments canvassed above don't seem to give functionalists any very strong reasons to doubt their doctrine. But there is a curious asymmetry in the arguments against functionalism. They exploit only one direction of the functionalist biconditional. Functionalists assert that each mental state is identified by its role within some functional architecture. So the claim is this: for every type of mental state, M, there is functional role, R, such that a system is in state M if and only if it has a functional architecture which implements R. The arguments against functionalism which we've seen, and which can be found in the literature, operate by finding something which possesses R but lacks M, contrary to the right to left direction of the functionalist biconditional.

Consider the classical inverted spectrum thought experiment. Here we are supposed to imagine two people who have inverted colour experience relative to each other: where the one sees green the other sees red, where the one sees yellow the other sees blue, and so on throughout the spectrum. The difficulty for functionalism is that it seems intuitively compelling, at first glance at least, that the two people will be behaviourally indistinguishable (they both learned to call the sun "yellow" no matter what they each experience when they glance at the sun). That is, there will be no behavioural evidence which rules out our hypothesis that they have inverted experiences. Nor does it seem that there has to be any hidden, internal difference in their cognitive architectures.

The inversion thought experiment has been attacked by denying these last points. Maybe there will be subtle behavioural differences between such people which will reveal a difference of functional role between the experience of red and green. One possibility: there are, I am told, just 18 steps of perceptually equal hue changes between the primary colours yellow and green, but 31 such steps between red and blue (see Hardin 1988, pp. 141-2). Maybe someone's inverted colour vision would be revealed by a simple test exploiting this fact. It has also been noted that colours have seemingly natural emotional affinities: red is aggressive, "hot", active, whereas blue is "cool", "down" and passive. People's moods can be and are intentionally modified by office and public room colour schemes. Again, maybe someone with inverted colour vision would have inappropriate emotional responses to such rooms. Such differences would be manifest reflections of more or less subtle functional differences in the inner states playing the proper roles of colour experiences. That is, it may be that the functional role of colour experiences is

rather wider and richer than philosophers normally imagine. (It is an interesting, if not entirely clear, question whether these "subsidiary" functional roles of colour experience are part of the ordinary concepts of visual consciousness.)

It is interesting to compare the inverted spectrum thought experiment with Siewert's blindsight thought experiments, now that we have these functionalist responses in mind. It is quite easy to believe that conscious visual experience might carry emotional affect that mere possession of the information provided by those experiences would not. We can try, by philosophical fiat, to impose upon the thought experiment the condition that all emotional consequences, at least insofar as these are capable of manifest exhibition, of both spontaneous blindsight and conscious vision be exactly the same. But maybe such a condition transgresses the bounds of nomological possibility. Consider the quotation from the deaf child above. Obviously there is great joy in simply being able to hear sounds (and I think I can also detect reference to a certain, rather different, kind of joy in the pure silence of her deafness as well). Part of why it seems so evident that conscious experience has intrinsic value is precisely the linkage - which seems deeper than mere cause and effect - between experience and more or less intensely felt positive emotional responses. Thus it may be that this linkage between conscious experience and emotions is, so to speak, internal to consciousness itself so that Connie will have a range of responses which will just never be available to Belinda. Now, so long as there is a functionalist account of emotional response, it will be possible to point to a functional difference between Belinda and Connie. We can now begin to replay the arguments against functionalism with regard to *emotional* rather than *visual* consciousness, but there is a palpable sense of spinning wheels here.

I want to conclude by making an effort to exploit the neglected left to right direction of the functionalist biconditional. That is, instead of trying to find two creatures with identical functional organization (whether at the manifest or hidden level) but different states of consciousness, let us try to find two creatures with distinct functional organization but identical states of consciousness. This would be just as damaging to functionalism as the traditional problems but has the advantage, as I see it, of being rather easier to establish, and this without recourse to any bizarre thought experiments.

Some preliminary remarks are necessary. The first is the introduction of the notion of a qualia-space. This is not unfamiliar. We are used to representing qualia, and especially colour qualia for which by far the most data are available, as forming a space ordered by a few basic parameters. For colour the general properties of hue, saturation and brightness can generate such a space as in the familiar Munsell colour solid reproduced (poorly and in cutaway form) here in Figure 1.

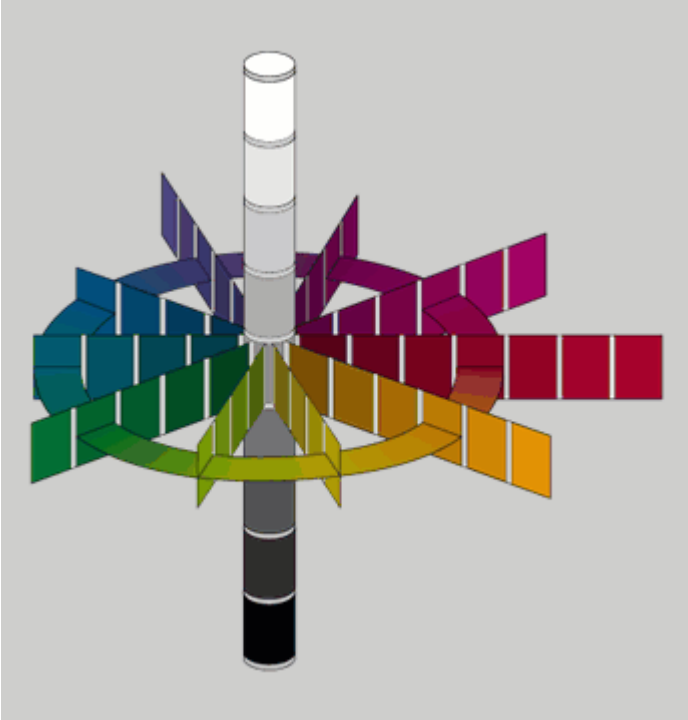

Figure 1.

Cutaway view of color solid.

Notice that functionalism has a natural affinity for this notion. In functionalism, mental states are *identified* by their "position" within a set of relations, relations which hold among other mental states, behavioural "outputs" and sensory "inputs". We might call the relations that hold amongst the mental states, and kinds of qualia in particular, intra-mental or intra-qualia relations, and the relations holding between mental states and behaviour and sensory input extra-mental or extra-qualia relations. Such a set of relations can be regarded as forming an abstract space and in fact the colour qualia space of the Munsell representation (if it does accurately represent phenomenological colour) would form a subspace, restricted to the intra-qualia relations, of the very much larger space specified by the full functional specification of colour qualia.

In terms of the idea of qualia spaces (or the more general notion of "mental state space" appropriate for the full functionalist treatment of the mind), many of the anti-functionalist thought experiments can be readily characterised. They are transformations of a qualia space which preserve its structure while transposing certain elements within it. That is a cryptic remark but an example will make things clearer. Consider again the Munsell colour solid. Notice that the hue circle is symmetrical. Thus we could map opposite hues, red and green for example, on to each other by "folding over" the hue circle. All the relations within the solid would be preserved by such a transformation - flipping or rotating the hue circle has no "relational effect". If one identifies colour qualia with positions in the Munsell solid (something I imagine few would actually agree to, since colour qualia possess even relational complexities that outrun the resources of this

representation, but that is irrelevant to the point I'm trying to make) then flipping or rotating the hue circle makes no difference whatsoever.

Now, if one believed that there were intrinsic, non-functionally specifiable, qualitative features of colour qualia then one might resist the idea that flipping the hue circle has no effect. In fact, one might think that the result would be *hue-qualia inversion.* (Other inversions would result if other symmetries are exploited. For example, brightness could be flipped leaving us with a thought experiment in which some people see "dark" where we see "light" and vice versa, even though there is no difference in their visual discriminatory abilities or, in general, in any of their visually mediated behaviour.) I believe that it is our intuitive sense that our colour qualia space has such symmetries within it that underwrites the appeal of the inverted spectrum objection to functionalism (see Campbell 2000). And it is only a small step from accepting qualia inversion to the claim that qualia are independent of functional organization and thus to the kind of thought experiments which Siewert favors.

The issue can be presented in this way. Functionalists regard mental states as identified by their location in an abstract space of relations, nodes in complex network if you will. Friends of qualia agree that qualia take up positions within such a structure but deny that their nature is thus exhausted. Qualia have an independent existence and causal powers which are in fact what allows them to play the roles the functionalists postulate. One can see a natural and possibly tempting compromise position here: let the qualia be the "stuff" which occupies the role in any particular system (see Churchland and Churchland 1982). This gives qualia an intrinsic non-relational nature without divorcing them from their appointed functional role. However, such a suggestion immediately runs foul of the obvious, and often mentioned, problem of chauvinism in that it entails that Martians (who, I say in good philosophical tradition, have hydraulic, liquid-methane based brains rather than our familiar electrical carbon based ones) necessarily have qualia different from ours. But this seems at least unmotivated. The Martians might give every sign of enjoying exactly the same qualia that we do, and the fact that they are made of different stuff doesn't seem to preclude this.

That is, it seems that qualia are no less multiply realizable than any other mental state. And yet multiple realizability is a doctrine that is naturally tied to functionalism and without this attendant functionalism threatens to be entirely inexplicable. *Roles* are the kind of thing that can be instantiated in a variety of realizing substrates, but when we get down to things possessing, or supposedly possessing, an intrinsic nature we arrive at something which cannot be multiply realized. The only consistent option for the lover of qualia is a non-physicalist account of them: both we and the Martians have physical stuff sufficient to *bring about* the very same qualia that we enjoy (or, more generally, their physical basis yields a "subjective perspective" identical to ours).

Thus despite his apparently agnostic attitude I think we can see that Siewert's views at least come very close to entailing the denial of physicalism (I'm not sure what he would say about the claim that qualia are "intrinsically physical" but it seems to me at odds with the tenor of his work ...).

Maybe the functionalist is happy now, since if we support the intuition that the Martian *might* have the same qualia as we do, despite vast physical differences, and refuse to abandon physicalism - and the functionalist will tend to assert that every right thinking philosopher must accept some form of physicalism - then it seems that we'll have to consider some more abstract identity between the Martians and us to underwrite the identity of qualitative consciousness. What could this abstract identity be save some (complex) relational property of the structure of both our own and the Martians' physical makeup? And this would amount to the endorsement of some form of functionalism. But does the idea of qualia "spaces" really support the functionalist?

Consider the following visual task (see Figure 2).

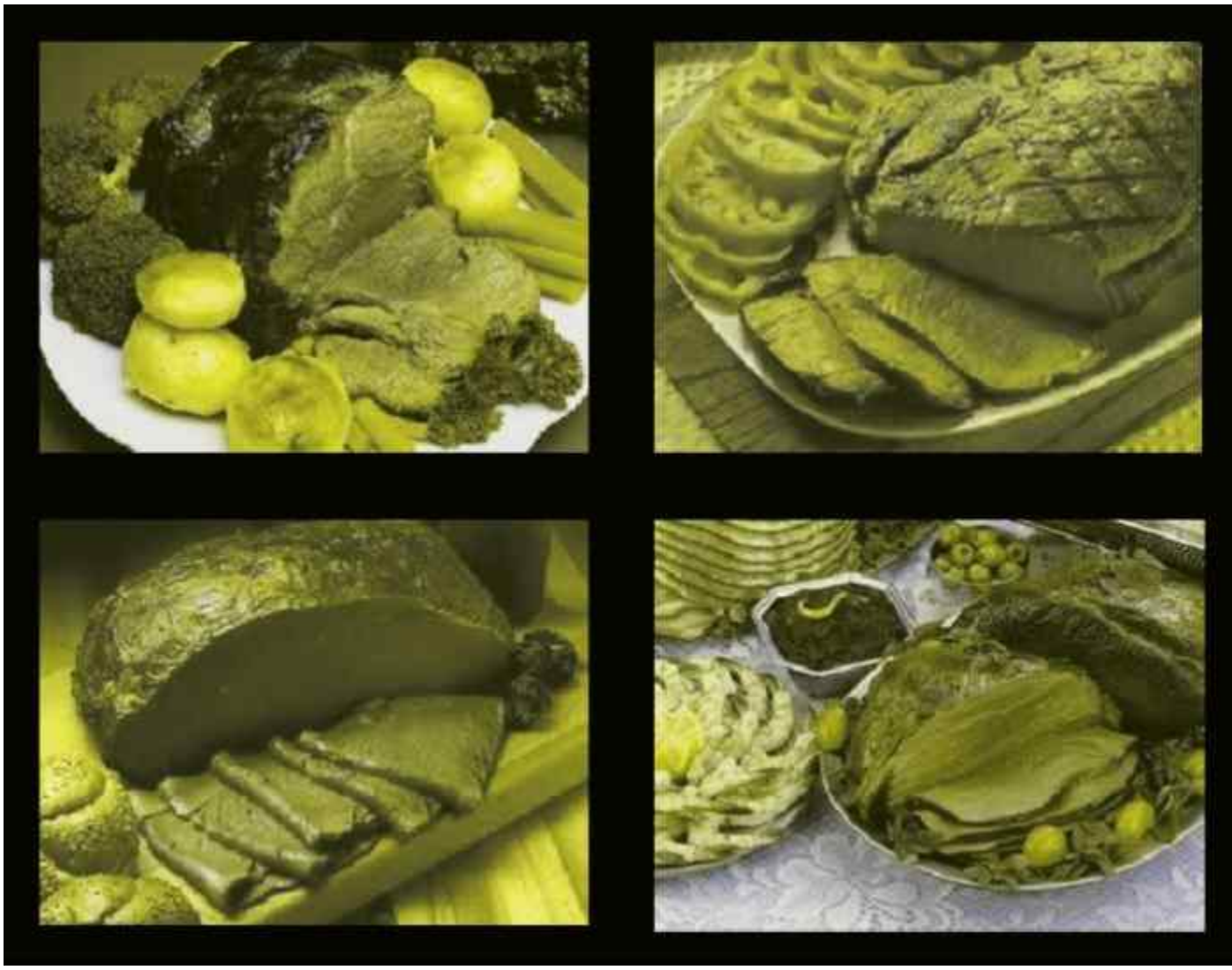

Figure 2.

Which is rarest?

Deciding which roast is the most rare is normally a simple job which obviously (normally) depends upon visual consciousness. But because the red-channel has been eliminated from this slide, it is basically impossible. In fact this slide is taken from a web page (http://www.mcw.edu/cellbio/colorvision) devoted to colour vision and its pathologies, and is a representation of how things look to someone suffering from deuteranopia, a kind of severe colour blindness which arises from loss of function of the M-cones, that is, the colour receptors whose peak sensitivity is between that of the other two types of cones.

It seems to me hard to deny that the slide could be an utterly faithful representation of how the world looks to a deuteranope (whether it is precisely faithful or not does not really matter to the point being made here). But notice that the scene is "made up of" the same colour qualia we all experience, although of course restricted to a subset of the full range of colour qualia normally sighted people can experience. The slide also makes it apparent why colour blind people fail to have the same manifest colour related dispositions as normally sighted people (such as the inability to distinguish reliably between a rare and a well done roast by eyesight alone). But the qualia available to the colour blind are not different qualia; they are the same qualia (or some of them) in a modified qualia-space (both intra- and extra-qualia relations have changed). In addition to the inherent plausibility of this sort of example, there is empirical evidence that the structurally different qualia spaces of the colour blind can support qualia identical to those of the colour sighted. Some people become colour blind as adults and can, seemingly, compare their qualia across the change. These people do not report a radical transformation of all their colours (Hardin reports that there are even some people who are colour blind in one eye, colour sighted in the other and they can directly compare the qualia arising from both eyes). The crucial point here is that although their colour vision is altered, what's left remains *colour* vision, made up of some of the same hues as possessed by those with normal colour vision.

However, this eminently plausible representation of what it is like to be colour blind would seem to be necessarily inaccurate given a functionalist account of colour qualia. To allow that the same qualia, as functionally definable entities, persist across changes in functional relationships is incoherent. The problem here is analogous to one in semantics: the problem of holism (see Fodor and Lepore 1992). The problem there is that if the meaning of a term, say, "vehicle" is determined by its place within a set of relationships to other terms, as well as relations to behaviour and sensory input (call this set a "semantic-space"), then any two people who differ in the way "vehicle" fits into their semantic-space will disagree about the meaning of "vehicle" (and every other word too of course). Now, we know that everyone will differ in some ways about the use of "vehicle" from just about everyone else. (I have seen this demonstrated in a large crowd via a series of questions: is a car a vehicle, is a train, a wagon, a horse, a scooter, etc. Soon it seemed that everyone in the room disagreed with everyone else about just what, *exactly*, a vehicle was.) Thus we end up with a radical semantic incommensurability and nobody ever understands what anybody else is talking about (this is the loneliness of the analytic philosopher).

Whether or not there is merit in the semantic argument, the case of qualia is strikingly worse, for we have no conception of a colour quale that, for example, is *like* green but which is not any shade of green, or any shade of any other colour that we can experience. That is, we cannot suppose that a deuteranope sees the roasts illustrated above as green, for we see them thus and we have a manifestly different functional architecture underlying our colour vision; nor can we suppose that the deuteranope sees the roasts as red, or purple, or grey or any other colour that *we* can experience, since all of these colours, as functionally definable entities, already have a place within *our* functional architecture. So the deuteranope ought to have a radically different sort of colour vision, of which we cannot even conceive. This seems very unlikely, and certainly does not jibe with what people who become colour blind say about their own visual consciousness.

A possible line of reply for the functionalist is to hypothesize that there must be some region of functional identity within the globally distinct architecture of colour vision. This would be to pick out some subset of the normal intra and extra qualia relations as *definitive* of, say, green qualia. (This line of reply is possible in the case of the semantical argument as well - it threatens to lead to the idea that there are "analytic connections" which define the meanings of words; this is not regarded as progress.) It would be reasonable to suppose that this subset is that involved with discrimination of green things and visual tasks restricted only to the domain of green things. Roughly speaking, the idea is that if the deuteranope and the normally sighted do equally well in environments which are colour restricted in the appropriate way then they can be said to share colour qualia thus restricted. Perhaps in a room in which everything is some shade of green, the normally sighted and colour blind are indistinguishable.

The difficulty now is that there are test situations, even whole rooms, with restricted colours but including *red* in which the colour blind and colour sighted will do equally well, but it is absurd to conclude from this that the colour blind can after all experience red. It is up to the functionalist about colour experience to describe the restricted qualia space of green which both fits into the larger space of full colour experience in a way that preserves the colour experience of the normally sighted while accounting for the visual deficits of the colour blind. The abstract problem is to specify the relevant functional architectures without begging the question by presupposing anything about the colour vision that is associated with various *distinct* architectures.

But solving that problem is the burden of functionalists.

References

Block, N. (1978). "Troubles With Functionalism", in W. Savage (Ed.) *Perception and Cognition: Minnesota Studies in the Philosophy of Science,* vol. 9, Minneapolis: University of Minnesota Press.

Campbell, N. (2000). "Revising the Inverted Spectrum", ms. (Presentation at the 2000 meetings of the Eastern Division of the APA.)

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press.

Churchland, P. and Churchland, P. (1982). "Functionalism, Qualia and Intentionality", in J. Biro and W. Shahan (eds.) *Mind, Brain and Function*, Norman: University of Oklahoma Press.

Fodor, J. and Ernest L. (1992). *Holism: A Shopper's Guide.* Oxford: Blackwell.

Hardin, C. L. (1988). *Color for Philosophers.* Indianapolis: Hackett.

Lem, S. (1980). "The Seventh Sally, or How Trurl's Own Perfection Led to No Good", in *The Cyberiad: Fables for the Cybernetic Age.* New York: Avon (Bard).

Rossetti, Y. (1998). "Implicit Perception in Action: Short-Lived Motor Representations of Space", in *Consciousness and Cognition*, Vol. 7, No. 3, pp. 520-558. (A pdf version of this paper can be found at http://www.lyon151.inserm.fr/unites.)

Searle, J. (1980). "Minds, Brains and Programs", *Behavioral and Brain Sciences, 3*, 417-24.

Siewert, C. (1998). *The Significance of Consciousness.* Princeton: Princeton University Press.