Reply to Mulhauser's Review of *The Conscious Mind*

David J. Chalmers
Department of Philosophy
University of California
Santa Cruz, CA 95064
USA

chalmers@paradox.ucsc.edu

PSYCHE, 2(35), November 1996 http://psyche.cs.monash.edu.au/v2/psyche-2-35-chalmers.html
KEYWORDS: consciousness, qualia, materialism, dualism, functionalism, computation, zombies, information.

RESPONSE TO: G. Mulhauser, 'Bridge Out' On The Road To A Theory Of Consciousness: A Review of *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, $29.95
hbk., pp. xvii + 414, ISBN 0-19-510553-2.

Gregory Mulhauser's vigorous review of my book *The Conscious Mind* includes quite a few misrepresentations and philosophical misunderstandings. I would not like these misunderstandings to be
perpetuated, so I will make a few comments in reply

First, a clarification on "taking consciousness seriously". Mulhauser characterizes this assumption as the assumption that no cognitive theory of consciousness will suffice. The latter assumption would indeed beg some crucial questions, but it is not the assumption that I make. I make an assumption about the *problem* of consciousness, not about any solution. To quote (p. xii):
Throughout the book, I have assumed that consciousness exists, and that to redefine the problem as that of explaining how certain cognitive and behavioral functions are performed is unacceptable. This is what I mean by taking consciousness seriously.
That is, the premise is simply that there is a phenomenon to be explained, and that the problems of explaining such functions as discrimination, integration, self-monitoring, reportability, and so on do not exhaust all the problems in the vicinity. The deepest problem of consciousness, as I understand it, is not the problem of how all these functions are performed, but rather the problem of explaining how and why all this activity supports states of subjective experience.

This isn't to make any assumptions about the nature of the solution -- plenty of people agree with the premise but still think that one way or another they can get a cognitive or materialist theory of consciousness to work -- although of course I do go on to *argue* that if the premise is granted, it turns out that such theories will always be incomplete.

Like many people (materialists and dualists alike), I find this premise simply obvious, although I can no more "prove" it than I can prove that I am conscious. At the very least, to deny this premise would require extraordinarily strong arguments, of a type that I have never seen. In my experience the large majority of people find it obvious; but certainly there are some that deny it, and arguments over whether the premise is true or false rapidly descend into table-pounding. Wishing to avoid too much of that dead end, I prefer to simply state the assumption up front. Mulhauser complains that the assumption is not "discharged", but of course that is the whole point of making it a premise. (I do argue for it where I can, but there is no denying that such arguments -- on either side -- ultimately come down to a bedrock of intuition at some point.) The result, as I also say up front, is that the minority of people who don't see a "hard problem" aren't going to find the book of more than intellectual interest. That's life, though some of them seem to read and review it anyway.

That, I think, is the deepest issue relevant to Mulhauser's review. In the rest of it I find mostly rhetoric, misunderstanding, and an impressive repertoire of adjectives. Actually, that's not quite fair. He makes a couple of good points on minor issues, although nothing that really affects the state of play.

The first point that looks like a substantial objection is in his discussion of Chapter 2, where he appears to have an objection to my notion of logical necessity and its relation to formal systems, but as he does not say what it is (he simply says that he has an objection), it is hard to evaluate. This pattern -- highly opinionated criticism without any substantiation -- is repeated many times in the course of the review.

His discussion of the central Chapter 3 mostly centers around the conceivability of zombies. I say explicitly in the text that this conceivability argument is more inconclusive than some of the others (due to the difficulty of assessing claims about conceivability), and that it does not bear the central burden of the argument, but commentators often focus on it anyway. In any case: Mulhauser's description of my zombies as "a particularly strong variety: a physical, functional, and *psychological* duplicate", with its emphasis, makes the claim seem stronger than it is: "psychological" is being used here in a stipulative sense that adds nothing over and above "functional". What is relevant to the arguments is that a zombie be a physical and functional duplicate; that's all.

Mulhauser's objection to the argument is the standard "how do we know we can really conceive this". He doesn't face up to the real point, which is the challenge to isolate an

inconsistency in the notion, and in particular to demonstrate the aspect of the concept of consciousness which could ground an conceptual entailment from the physical facts to the facts about consciousness. He neither comes out and embraces a functional definition of consciousness (equating what needs to be explained with reportability, integration, or some such complex function), which would require facing up to all the problems with such a definition, nor does he even hint at another way that such an entailment might go through, which would require facing up to the fact (stressed in the book) that structure and function only ever adds up to more structure and function. So we are left in a position of assertion without enlightenment.

His discussion of epistemic asymmetry makes his first good point: that a logical entailment from physical facts to facts about consciousness requires a grasp of the concept of consciousness, so the mere fact that someone without consciousness wouldn't derive facts about consciousness from the physical facts can be explained away by the observation that they lack the concept. This is a minor point -- there is still a striking disanalogy here between our epistemic access to consciousness and other phenomena -- but nevertheless I should have rephrased things in a way that brings that disanalogy out more directly. In any case, I also point out the epistemic asymmetry of consciousness in ways that aren't vulnerable to this objection, e.g. by giving the observer the concept of consciousness and noting that the facts about consciousness in others remain underdetermined.

His second interesting point is the idea that the explanatory gap may arise for reasons of formal undecidability or computational intractability. There are obvious problems to be addressed: either consciousness is a structural/functional concept, in which case (a) it is most unclear why the facts about it should be any more undecidable or intractable than the facts about any other such concept (learning, reportability, or whatever), and (b) the view will presumably be vulnerable to arguments just like mine, establishing a gap between the structural/functional concept and consciousness while leaving the "intractable" physical details to one side; or it is not a structural/functional concept, in which case there are systematic reasons why it cannot be logically necessitated by the physical facts, intractability/undecidability or no. So I think even this sort of position is vulnerable to the sort of critique I provide. But in any case it would be very interesting to see such a view worked out in detail, and the project of doing so would be sure to be enlightening. I do address this view briefly on my pp. 138-40 -- not in much depth, as the view has not been explicitly set out in the literature as far as I know, so I had to more or less make it up -- but there is no doubt that there is more to say about it, and it would be a very useful service to see the view worked out systematically.

Mulhauser's discussion of "explanatory exclusion" is off the point. His claim that my intuitions about consciousness are based in "explanatory exclusion" intuitions that apply equally to planets, dolphins, and so on ignores the many pages of argument earlier in the book analyzing the disanalogy in depth. Indeed, none of the arguments I give about

consciousness would have a hope of getting off the ground when applied to planets, dolphins, and the like.

His discussion of the epistemological problems of consciousness adds nothing of substance. There are, of course, serious problems here -- the first-person epistemology of consciousness poses some of the deepest questions in the field, and questions to which I have at best tentative and incomplete answers. But Mulhauser gives no reason to believe that these answers must be bad ones; he simply reiterates the problems, as I have laid them out. He is of course absolutely right that many of these problems can be avoided if we embrace a reductive functionalist view of consciousness (where all that needs to be explained is discrimination, integration, reportability, and other complex functions). In fact, if we embrace such a view, almost *all* the problems of consciousness are removed! For those who think this sort of solution-by-stipulation provides any enlightenment, they are welcome to it.

The entire second half of Mulhauser's review is thrown deeply astray by his "discovery" of my "little secret" that I think functional awareness is logically necessary for consciousness: a "discovery" based solely on my statement "awareness is necessary for consciousness". But in context it is obvious that natural necessity is at issue -- what is under discussion are the bridging laws connecting processing and consciousness -- and the sentence at the end of the paragraph makes it clear that logical necessity couldn't possibly be what is meant. Indeed, this "discovery" is so much at odds with everything else in the book that its invocation suggests that the reviewer is holding a principle of charity far in abeyance.

Mulhauser also seems to think that my suggestions for psychophysical bridging principles are *a priori* deductions from the single fact that consciousness exists. This is simply false: they are inferences drawn from apparent regularities between facts about functioning and facts about experience, especially in the first-person case. I start by drawing connections between (1) functional facts about reportability, behavioral control, and the like, and (2) phenomenological observations about the structure and properties of experience in the first-person case. These connections are certainly quite coarse (I make no claim to be a sophisticated phenomenologist) and I would be all for a project of refining them. But in all of them phenomenological data -- observations about the structure of experience, for example, or its presence or absence in different sorts of cases -- play a significant role. In every case the relevant data go well beyond "consciousness exists"; Mulhauser is quite right that no substantial inferences about the nature of psychophysical laws could be drawn from that datum alone.

(Some may object to using phenomenological data, but I would argue that without an appeal to such data, a theory of conscious experience as such cannot even get off the ground -- it will be a third-person theory that never connects to the first-person. And indeed I would argue that everyone who draws conclusions about experience from

processing data is implicitly relying on bridging principles that are partly based on first-person knowledge of phenomenology; I am just trying to bring them into the open.)

Mulhauser's discussion of the fading and dancing qualia arguments makes a number of mistakes. The arguments cannot come close to establishing that absent or inverted qualia are logically impossible, for reasons I address directly: first, nothing in the arguments suggests that fading and dancing qualia scenarios are incoherent, and second, they all take the empirical fact of the existence of consciousness as a premise. Mulhauser somehow slides from this second observation of mine to the reiterated "conclusion" that my psychophysical principles are all deduced a priori from this observation, but this is an elementary mistake -- to say that an argument takes something as a premise is not to say that it takes it as its *only* premise! Mulhauser also makes a lot of play out of a tension between my arguments here and my supposed view that awareness is logically necessary for consciousness; but as noted above, this view is a product of his imagination.

For what it's worth, Mulhauser's suggestion that this supposed view implies that dancing qualia are logically impossible seems to commit another elementary philosophical mistake: someone could believe that awareness is logically necessary for consciousness but not logically sufficient, and in particular could believe that facts about awareness do not logically entail facts about the intrinsic character of specific qualia.

Mulhauser's discussion of the last three chapters of the book is vigorous but largely free of substance. He complains that I don't appeal to his favorite sort of information theory, namely Chaitin's. I think Chaitin's work is fascinating (certainly it is intrinsically much more interesting than Shannon's), but it is an entirely different creature and I didn't see its application here. If Mulhauser can find a way to exploit it in a theory of consciousness, more power to him. His further observations (e.g., that some sorts of information processing seem to happen without consciousness) simply go over ground already covered in the book, without addressing my discussion.

His complaints about my chapter on strong AI rest on my "failure to grasp the vacuity of naive functional theorizing", and on his appeal to philosophers such as Putnam who have pointed out this "vacuity". Like many, I am not deeply impressed by these arguments, although they have some interest. I have a lengthy analysis of the Putnam considerations in my "Does a Rock Implement Every Finite State Automaton?" (*Synthese* 108:309-33, 1996), and the general issues are discussed in "A Computational Foundation for the Study of Cognition&quot (of which a shorter version appeared in *Minds and Machines*, 1994). Mulhauser provides nothing more than an appeal to authority here, though, so there is little I can say in reply. And although Mulhauser suggests that a giant look-up table might be constructed to satisfy my CSA definition, I would be very interested to see such a CSA that is naturally possible (remember, it is natural possibility that is relevant here). As I discuss in my paper on Putnam's argument, the only such "false implementations" are probably beyond the bounds of natural possibility.

Mulhauser says nothing of substance in objecting to my discussion of incompleteness -- once again, he simply spends a long paragraph asserting vigorously that he has objections. The same for his discussion of discreteness vs. continuity, in the next paragraph. Those who understand the issues will know that the Siegelmann/Sontag machine he refers to does not threaten anything I say: like every other "super-Turing" machine that has been proposed, it requires an infinite resource or infinitely precise initial conditions in order to compute a super-Turing number-theoretic function. With bounded precision, or even with unbounded precision but a fixed finite initial state, it does no better than a Turing machine. As with many other places in the review, the argument strategy seems to be "unargued appeal to irrelevant authority", which ends up merely betraying the reviewer's lack of understanding.

In his discussion of my analysis of the Everett interpretation of quantum mechanics, Mulhauser asserts quite correctly that (1) a similar analysis might be available to a materialist (assuming a materialist theory of consciousness could work at all), and (2) that these considerations do not provide knockdown support for the interpretation (as I say in the book, they simply remove one sort of objection, and thus give it indirect support; the interpretation still has serious problems, and indeed I am not sure whether I believe it). He apparently intends these remarks as criticisms, but it is hard to see why. The chapter is not intended as religious warfare either for this interpretation or for dualism; it is simply an exploration of the issues.

I will end with a few comments on the more unusual aspects of Mulhauser's review. Mulhauser seems much concerned to point out everywhere that an argument I use has appeared in the literature before. I am not sure why, as I acknowledge my multiple debts right up front, and I cite every debt as the book goes along (it's not for nothing that the book has a 400-entry bibliography). Apparently he is spurred on by "the widespread publicity...suggesting it would contain startling new arguments and thought experiments". He would not be the first reviewer to review the media rather than the book, but it would be nice to have the book reviewed for what it is.

In reality, it's a book written by a mere mortal, and a book which tries to present the arguments -- some old, some new -- as carefully and as completely as possible. There's no question that anti-reductionist arguments have been around a long time, and negative arguments with the same general thrust as mine have been seen before, though they have usually been presented in a very piecemeal way. In the book, I try to really make the case properly: maybe I succeed and maybe I fail, but I make no apologies for retreading old ground at those points where I think it is the best way to do things.

I must respond to one particularly odd remark of Mulhauser's in this context. A characterization of my fading and dancing qualia arguments -- "both of which, contrary to popular tale, have appeared in one form or another elsewhere in the literature" -- seems

to suggest that I have borrowed someone else's arguments unacknowledged. In fact both arguments are accompanied by citations to every item I know in the literature to which they bear even a remote similarity. The fading qualia gradual-replacement scenario is of course an old folk tale in the literature, and in the paragraph where it is introduced I cite a number of papers that use it (although they use it in somewhat different ways), and I footnote some more. The dancing qualia argument, as far as I know, was original with me in my 1993 Ph.D. dissertation, although there do exist a couple of distantly related scenarios used to argue for quite different conclusions, and a more closely related version was formulated independently by Arnold Zuboff in a 1994 article. So as with much of the rest of the book, there are a number of elements which are familiar, and a number which are new. That's the way it usually is in philosophy.

Finally, let me express my bemusement at the religious tone that is so common in this sort of discussion. I don't know why discussions of consciousness provoke this tone, which is not nearly so pronounced in other areas of science and philosophy. I don't feel any particular emotional commitment to my conclusions myself (I may be a materialist in my heart, if not in my mind); I have come to the conclusion that they more or less have to be true, for systematic reasons, but if someone thinks they have a better option, that's fine. I tried to write a book setting out the lay of the land as I see it, laying out what seemed to me to be a promising approach, and openly acknowledging the various problems my approach faces along the way. If someone disagrees, then I'm interested to hear about it. It's a fascinating area of intellectual inquiry. But cheap point-scoring and unargued rhetoric are uninteresting. Mulhauser does make a couple of points of substance along the way; I suggest that we are all best served if substance is what we stick to.

References

Chaitin, G.J. (1990). *Information, randomness, and incompleteness.* (2nd ed.). World Scientific.

Chalmers, D.J. (1994). A computational foundation for the study of cognition. *PNP Technical Report 94-03*, Washington University. [http://ling.ucsc.edu/~chalmers/papers/computation.html]. Shorter version published as "On implementing a computation". *Minds and Machines, 4,* 391-402.

Chalmers, D.J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press. [http://ling.ucsc.edu/~chalmers/tcm.html]

Chalmers, D.J. 1996. Does a rock implement every finite-state automaton? *Synthese, 108,* 309-33. [http://ling.ucsc.edu/~chalmers/paper/rock.html]

Putnam, H. (1988). *Representation and reality.* MIT Press.

Siegelmann, H.T. & Sontag, E.D. (1994). Analog computation via neural networks. *Theoretical Computer Science, 131,* 331-60.

Zuboff, A. (1994). What is a mind? In P. French, T. Uehling, & H. Wettstein (Eds.),*Philosophical Naturalism.* University of Notre Dame Press.