On the Search for the Neural Correlate of Consciousness

David J. Chalmers
Philosophy Program
Research School of Social Sciences
Australian National University

I will discuss one aspect of the role that neuroscience plays in the search for a theory of consciousness. Whether or not neuroscience can solve all the problems of consciousness singlehandedly, it undoubtedly has a major role to play. Recent years have seen striking progress in neurobiological research bearing on the problems of consciousness. The conceptual foundations of this sort of research, however, are only beginning to be laid. I will look at some of the things that are going on, from a philosopher's perspective, and will try to say something helpful about these foundations.

We have all been hearing a lot about the "neural correlate of consciousness". This phrase is intended to refer to the neural system or systems primarily associated with conscious experience. The acronym of the day is "NCC". We all have an NCC inside our head, we just have to find out what it is. In recent years we have seen quite a few proposals about the identity of the NCC. One of the most famous proposals is Crick and Koch's suggestion concerning 40-hertz oscillations. That proposal has since faded away a little but there are all sorts of other suggestions out there. The picture is almost reminiscent of particle physics, where they have something like 236 particles and people talk about the "particle zoo". In studying consciousness, one might talk about the "neural correlate zoo".

A brief list of suggestions that have been put forward includes:

- 40-hertz oscillations in the cerebral cortex (Crick and Koch 1990).
- Intralaminar nucleus in the thalamus (Bogen 1995).
- Re-entrant loops in thalamocortical systems (Edelman 1989).
- 40-hertz rhythmic activity in thalamocortical systems (Llinas et al 1994).
- Nucleus reticularis (Taylor and Alavi 1995).

This paper appears in *Toward a Science of Consciousness II: The Second Tucson Discussions and Debates* (S. Hameroff, A. Kaszniak, and A.Scott, eds), published with MIT Press in 1998. It is a transcript of my talk at the second Tucson conference in April 1996, lightly edited to include the contents of overheads and to exclude some diversions with a consciousness meter. A more in-depth argument for some of the claims in this paper can be found in Chapter 6 of my book *The Conscious Mind* (Chalmers, 1996).

- Extended reticular-thalamic activation system (Newman and Baars 1993).
- Anterior cingulate system (Cotterill 1994).
- Neural assemblies bound by NMDA (Flohr 1995).
- Temporally-extended neural activity (Libet 1994).
- Back-projections to lower cortical areas (Cauller and Kulics 1991).
- Visual processing within the ventral stream (Milner and Goodale 1995).
- Neurons in visual cortex projecting to prefrontal areas (Crick and Koch 1995).
- Neural activity in area V5 (Tootell *et al* 1995).
- Certain neurons in the superior temporal sulcus and inferior temporal cortex (Logothetis and Schall 1989, Sheinberg and Logothetis 1997).
- Neuronal gestalts in an epicenter (Greenfield 1995).
- Outputs of a comparator system in the hippocampus (Gray 1995).
- Quantum coherence in microtubules (Hameroff 1994).
- Global workspace (Baars 1988).
- High-quality representations (Farah 1994).
- Selector inputs to action systems (Shallice 1988).

The list includes a few "cognitive correlates" of consciousness (CCC?) but the general idea is similar. We find some intriguing commonalities among the proposals in this list. A number of them give a major role to interactions between the thalamus and the cortex, for example. All the same, the great number and diversity of the proposals can be overwhelming. I propose to step back and try to make sense of all this activity by asking some foundational questions.

A primary question is this: how *can* one search for the neural correlate of consciousness? As we all know, measuring consciousness is problematic. The phenomenon is not directly and straightforwardly observable. It would be much easier if we had a way of getting at consciousness directly—if we had, for example, a consciousness meter.

If we had such an instrument, searching for the NCC would be straightforward. We would wave the consciousness meter and measure a subject's consciousness directly. At the same time, we would monitor the underlying brain processes. After a number of trials, we would say such-and-such brain processes are correlated with experiences of various kinds, so there is the neural correlate of consciousness.

Alas, we have no consciousness meter, and for principled reasons it seems we cannot have one. Consciousness just is not the sort of thing that can be measured directly. What, then, do we do without a consciousness meter? How can the search go forward? How does all this experimental research proceed?

I think the answer is this: we get there with principles of *interpretation*, by which we interpret physical systems to judge the presence of consciousness. We might call these *pre-experimental bridging principles*. They are the criteria that we bring to bear in looking at systems to say (1) whether or not they are conscious now, and (2) which information they are conscious of, and which they are not. We cannot reach in directly and grab those experiences, so we rely on external criteria instead.

That is a perfectly reasonable thing to do. But something interesting is going on. These principles of interpretation are not themselves experimentally determined or experimentally tested. In a sense they are pre-experimental assumptions. Experimental research gives us a lot of information about processing; then we bring in the bridging principles to interpret the experimental results, whatever those results may be. They are the principles by which we make *inferences* from facts about processing to facts about consciousness, and so they are conceptually prior to the experiments themselves. We cannot actually refine them experimentally (except perhaps by first-person experimentation!), because we have no independent access to the independent variable. Instead, these principles will be based on some combination of (1) conceptual judgments about what counts as a conscious process, and (2) information gleaned from our first-person perspective on our own consciousness.

I think we are all stuck in this boat. The point applies whether one is a reductionist or an anti-reductionist about consciousness. A hard-line reductionist might put some of these points a little differently, but either way the experimental work will require pre-experimental reasoning to determine the criteria for ascribing consciousness. Of course such principles are usually left implicit in empirical research. We do not usually see papers saying "Here is the bridging principle, here are the data, and here is what follows." But it is useful to make them explicit. The very presence of these principles has strong and interesting consequences in the search for the NCC.

In a sense, in relying on these principles we are taking a leap into the epistemological unknown. Because we do not measure consciousness directly, we have to make something of a leap of faith. It may not be a big leap, but nevertheless it suggests that everyone doing this work is engaged in philosophical reasoning. Of course one can always choose to stay on solid ground, talking about the empirical results in a neutral way, but the price of doing so is that one gains no particular insight into consciousness. Conversely, as soon as we draw any conclusions about consciousness, we have gone beyond the information given. So we need to pay careful attention to the reasoning involved.

What are these principles of interpretation? The first and by far the most prevalent is the principle of verbal report. When someone says "Yes, I see that table now", we infer that they

are conscious of the table. When someone says "Yes, I see red now", we infer that they are having an experience of red. Of course one might always say "How do you know?"—a philosopher might suggest that we may be faced with a fully functioning zombie—but in fact most of us do not believe that the people around us are zombies, and in practice we are quite prepared to rely on this principle. As pre-experimental assumptions go, this one is relatively safe—it does not require a huge leap of faith—and it is very widely used.

The principle here is that when information is verbally reported, it is conscious. One can extend this principle slightly, for no one believes an *actual* verbal report is required for consciousness; we are conscious of much more than we report on any given occasion. Thus an extended principle might say that when information is directly *available* for verbal report, it is conscious.

Experimental researchers do not rely only on these principles of verbal report and reportability. The principles can be somewhat limiting when we want to do broader experiments. In particular, we do not want to restrict our studies of consciousness to subjects that have language. In fact, at this conference we saw a beautiful example of research on consciousness in language-free creatures. I refer to the work by Nikos Logothetis and his colleagues (e.g., Logothetis and Schall 1989; Leopold and Logothetis 1996, Sheinberg and Logothetis 1997). This work uses experiments on monkeys to draw conclusions about the neural processes associated with consciousness. How do Logothetis *et al* manage to draw conclusions about a monkey's consciousness without getting any verbal reports? They rely on the monkey pressing bars: if the monkey can be made to press a bar in an appropriate way in response to a stimulus, we can say that that stimulus was consciously perceived.

The criterion at play seems to be require that the information be available for an arbitrary response. If it turned out that the monkey could press a bar in response to a red light but could do nothing else, we would be tempted to say that it was not a case of consciousness at all, but some sort of subconscious connection. If on the other hand we find information that is available for response in all sorts of different ways, then we will say that it is conscious.

The underlying general principle is something like this: When information is *directly available for global control* in a cognitive system, then it is conscious. If information is available for response in many motor modalities, we will say it is conscious, at least in a range of relatively familiar systems such as humans, other primates and so on. This principle squares well with the preceding one, when the capacity for verbal report is present: availability for verbal report and availability for global control seem to go together in such cases (report is one of the key aspects of control, after all, and it is rare to find information

that is reportable but not available more widely). But this principle is also applicable when language is not present.

A correlation between consciousness and global availability (for short) seems to fit the first-person evidence—that gleaned from our own conscious experience—quite well. When information is present in my consciousness, it is generally reportable, and it can generally be brought to bear in controlling behavior in all sorts of ways. I can talk about it, I can point in the general direction of a stimulus, I can press bars, and so on. Conversely, when we find information that is directly available in this way for report and other aspects of control, it is generally conscious information. One can bear out this idea by considering cases.

There are some tricky puzzle cases to consider, such as blindsight, where one has *some* availability for control but arguably no conscious experience. Those cases might best be handled by invoking the directness criterion: insofar as the information here is available for report and other control processes at all, the availability is indirect by comparison to the direct and automatic availability in standard cases. One might also stipulate that it is availability for *voluntary* control that is relevant, to deal with cases of involuntary unconscious response, although that is a complex issue. I discuss a number of puzzle cases in more detail elsewhere (Chalmers 1997), where I also give a much more detailed defense of the idea that something like global availability is the key pre-empirical criterion for the ascription of consciousness.

But this principle remains at best a first-order approximation of the functional criteria that come into play. I am less interested today in getting all the fine details right than in exploring the consequences of the idea that some such functional criterion is required and is implicit in all the empirical research on the neural correlate of consciousness. If you disagree with the criterion I have suggested—presumably because you can think of counterexamples—you may want to use those examples to refine it or to come up with a better criterion of your own. The crucial point is that in the very act of experimentally distinguishing conscious from unconscious processes, *some* such criterion is always at play.

My question is: If something like this is right, then what follows? That is, if some such bridging principles are implicit in the methodology for the search for the NCC, then what are the consequences? I will use global availability as my main functional criterion in this discussion, but many of the points should generalize.

The first thing one can do is produce what philosophers might call a *rational reconstruction* of the search for the neural correlate of consciousness. With a rational reconstruction we can say: Maybe things do not work exactly like this in practice, but the rational underpinnings of the procedure have something like this form. That is, if one were to try to *justify* the conclusions one has reached as well as possible, one's justification would

follow the shape of the rational reconstruction. In this case, a rational reconstruction might look something like this:

(1) Consciousness <-> global availability (bridging principle)

(2) Global availability <-> neural process N (empirical work)

so

(3) Consciousness <-> neural process N (conclusion).

According to this reconstruction, one implicitly embraces some sort of pre-experimental bridging principle that one finds plausible on independent grounds, such as conceptual or phenomenological grounds. Then one does the empirical research. Instead of measuring consciousness directly, we detect the functional property. We see that when this functional property (e.g., global availability) is present, it is correlated with a specific neural process (e.g., 40-hertz oscillations). Combining the pre-empirical premise and the empirical result, we arrive at the conclusion that this neural process is a candidate for the NCC.

Of course it does not work nearly so simply in practice. The two stages are highly intertwined; our pre-experimental principles may themselves be refined as experimental research goes along. Nevertheless I think one can make a separation into pre-empirical and experimental components for the sake of analysis. With this rational reconstruction in hand, what sort of conclusions follow? I want to draw out about six consequences here.

(1) The first conclusion is a characterization of the neural correlates of consciousness. If the NCC is arrived at via this methodology, then whatever it turns out to be, it will be a *mechanism of global availability*. The presence of the NCC wherever global availability is present suggests that it is a mechanism that *subserves* global availability in the brain. The only alternative is that it might be a *symptom* rather than a *mechanism* of global availability; but in principle that possibility ought to be addressable by dissociation studies, lesioning, and so on. If a process is a mere symptom of availability, we ought to be able to empirically dissociate it from global availability while leaving the latter intact. The resulting data would suggest to us that consciousness can be present even when the neural process in question is not, thus indicating that it was not a perfect correlate of consciousness after all.

(A related line of reasoning supports the idea that a true NCC must be a mechanism of *direct* availability for global control. In principle, mechanisms of indirect availability will dissociable from the empirical evidence for consciousness, for example by directly stimulating the mechanisms of direct availability. The indirect mechanisms will be “screened

off" by the direct mechanisms in much the same way as the retina is screened off as an NCC by the visual cortex.)

In fact, if one looks at the various proposals, this template seems to fit them pretty well. For example, the 40-hertz oscillations discussed by Crick and Koch were put forward precisely because of the role they might have in binding and integrating information into working memory, and working memory is of course a major mechanism whereby information is made available for global control in a cognitive system. Similarly, it is plausible that Libet's extended neural activity is relevant precisely because the temporal extendedness of activity gives certain information the capacity to dominate later processes that lead to control. Baars' global workspace is a particularly explicit example of such a mechanism; it is put forward explicitly as a mechanism whereby information can be globally disseminated. All these mechanisms and many of the others seem to be candidates for mechanisms of global availability in the brain.

(2) This reconstruction suggests that a full story about the neural processes associated with consciousness will do two things. First, it will *explain* global availability in the brain. Once we know all about the relevant neural processes, we will know precisely how information is made directly available for global control in the brain, and this will be an explanation in the full sense. Global availability is a functional property, and as always the problem of explaining the a function's performance is a problem to which mechanistic explanation is well-suited. So we can be confident that in a century or two global availability will be straightforwardly explained. Second, this explanation of availability will do something else: It will isolate the processes that *underlie* consciousness itself. If the bridging principle is granted, then mechanisms of availability will automatically be correlates of phenomenology in the full sense.

Now, I do not think this is a full *explanation* for consciousness. One can always ask why these processes of availability should give rise to consciousness in the first place. As yet we cannot explain why they do so, and it may well be that full details about the processes of availability will still fail to answer this question. Certainly, nothing in the standard methodology I have outlined answers the question; that methodology *assumes* a relation between availability and consciousness, and therefore does nothing to *explain* it. The relationship between the two is instead taken as something of a primitive. So the hard problem remains. But who knows: Somewhere along the line we may be led to the relevant insights that show why the link is there, and the hard problem may then be solved. In the meantime, whether or not we have solved the hard problem, we may nevertheless have isolated the *basis*

for consciousness in the brain. We just have to keep in mind the distinction between correlation and explanation.

(3) Given this paradigm, it is likely that there will be many neural correlates of consciousness. This suggestion is unsurprising, but the rational reconstruction illustrates just why such a multiplicity of correlates should exist. There will be many neural correlates of consciousness because there will be many mechanisms of global availability. There will be mechanisms in different modalities: the mechanisms of visual availability may be quite different from the mechanisms of auditory availability, for example. (Of course they *may* be the same, in that we could find a later area that would integrate and disseminate all this information, but that is an open question.) There will also be mechanisms at different stages in the processing path whereby information is made globally available: early mechanisms and later ones. All these may be candidates for the NCC. And we will find mechanisms at many levels of description: for example, 40-hertz oscillations may well be redescribed as high-quality representations, or as part of a global workspace, at a different level of description. It may therefore turn out that a number of the animals in the zoo, so to speak, can co-exist because they are compatible in one of these ways.

I will not speculate much further on just what the neural correlates of consciousness *are*. No doubt some of the ideas in the initial list will prove to be entirely off the track, and some of the others will prove closer to the mark. But I hope the conceptual issues are becoming clearer.

(4) This way of thinking about things allows one to make sense of a idea that is sometimes floated: that of a *consciousness module*. Sometimes this notion is disparaged; sometimes it is embraced. But the methodology in the search for an NCC suggests that it is at least possible that there could turn out to be such a module. What would it take? It would require some sort of functionally localizable, internally integrated area, through which all global availability runs. It need not be anatomically localizable, but to qualify as a module it would need to be localizable in some broader sense. For example, the parts of the module would have to have high-bandwidth communication among themselves, compared to the relatively low-bandwidth communication that they have with other areas. Such a thing *could* turn out to exist. It does not strike me as especially likely that things will turn out this way; it seems just as likely that we will find multiple independent mechanisms of global availability in the brain, scattered around without much mutual integration. If that is the result, we will probably say that there is no consciousness module after all. But that is another of those empirical questions.

If something like this module did turn out to exist in the brain, it would resemble Baars' conception of a global workspace: a functional area responsible for integrating information in the brain and disseminating it to multiple nonconscious specialized processes. In fact, many of the ideas I put forward here are compatible with things that Baars has been saying for years about the role of global availability in the study of consciousness. Indeed, this way of looking at things suggests that some of his ideas are almost forced on us by the methodology. The special epistemological role of global availability helps explain why the idea of a global workspace is a useful way of thinking about almost any empirical proposal about consciousness. If NCCs are identified as such precisely because of their role in global control, then at least on a first approximation, we should expect the global workspace idea to be a natural fit.

(5) We can also apply this picture to a question that has been discussed frequently at this conference: Are the neural correlates of *visual* consciousness to be found in V1, in the extrastriate visual cortex, or elsewhere? If our picture of the methodology is correct, then the answer presumably will depend on which visual area is most directly implicated in global availability.

Crick and Koch have suggested that the visual NCC is not to be found within V1, because V1 does not contain neurons that project to the prefrontal cortex. This reasoning has been criticized by Ned Block for conflating access consciousness and phenomenal consciousness (see Block, this volume); but interestingly, the picture I have developed suggests that it may be good reasoning. The prefrontal cortex is known to be associated with control processes; so *if* a given area in the visual cortex projects to prefrontal areas, then it may well be a mechanism of direct availability. And if it does not project in this way, it is less likely to be such a mechanism; at best it might be *indirectly* associated with global availability. Of course we still have plenty of room to raise questions about the empirical details. But the broader point is that for the sort of reasons discussed in (2) above, it is likely that the neural processes involved in *explaining* access consciousness will simultaneously be involved in a story about the *basis* of phenomenal consciousness. If something like this idea is implicit in their reasoning, Crick and Koch might escape the charge of conflation. Of course the reasoning depends on these somewhat shaky bridging principles, but then all work on the neural correlates of consciousness must appeal to such principles somewhere, so this limitation cannont be held against Crick and Koch in particular.

(6) Sometimes the neural correlate of consciousness is conceived of as the Holy Grail for a theory of consciousness. It will make everything fall into place. For example, once we

discover the NCC, then we will have a definitive test for consciousness, enabling us to discover consciousness wherever it arises. That is, we might use the neural correlate itself as a sort of consciousness meter. If a system has 40-hertz oscillations (say), then it is conscious; if it has none, then it is not conscious. Or if a thalamocortical system turns out to be the NCC, then a system without such a system is unlikely to be conscious. This sort of reasoning is not usually put quite so baldly, but I think one finds some version of it quite frequently.

This reasoning can be tempting, but one should not succumb to the temptation. Given the very methodology that comes into play here, we have no way of definitely establishing a given NCC as an independent test for consciousness. The primary criterion for consciousness will always remain the functional property we started with: global availability, or verbal report, or whatever. That is how we discovered the correlations in the first place. 40-hertz oscillations (or whatever) are relevant *only* because of their role in satisfying this criterion. True, in cases where we know that this association between the NCC and the functional property is present, the NCC might itself function as a sort of "signature" of consciousness, but once we dissociate the NCC from the functional property, all bets are off. To take an extreme example: If we have 40-hertz oscillations in a test tube, that condition almost certainly will not yield consciousness. But the point applies equally in less extreme cases. Because it was the bridging principles that gave us all the traction in the search for an NCC in the first place, it is not clear that anything follows in cases where the functional criterion is thrown it away. So there is no free lunch here: one cannot get something for nothing.

Once we recognize the central role of pre-experimental assumptions in the search for the NCC, we realize that there are limitations on just what we can expect this search to tell us. Still, whether or not the NCC is the Holy Grail, I hope that I have said enough to make it clear that the quest for it is likely to enhance our understanding considerably. And I hope to have made a case that philosophy and neuroscience can come together to help clarify some of the deep problems involved in the study of consciousness.

Bibliography

Baars, B.J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Bogen, J.E. 1995. On the neurophysiology of consciousness, parts I and II. *Consciousness and Cognition*, 4:52-62 & 4:137-58.

Cauller, L.J. & Kulics, A.T. 1991. The neural basis of the behaviorally relevant N1 component of the somatosensory evoked potential in awake monkeys: Evidence that backward cortical projections signal conscious touch sensation. *Experimental Brain Research* 84:607-619.

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

Chalmers, D.J. 1997. Availability: the cognitive basis of experience? *Behavioral and Brain Sciences* 20: 148-9. Also in N. Block, O. Flanagan, & G. Güzeldere (eds) *The Nature of Consciousness* (MIT Press, 1997).

Cotterill, R. 1994. On the unity of conscious experience. *Journal of Consciousness Studies* 2: 290-311.

Crick, F. and Koch, C. 1990. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2: 263-275.

Crick, F. & Koch, C. 1995. Are we aware of neural activity in primary visual cortex? *Nature* 375: 121-23.

Edelman, G.M. 1989. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

Farah, M.J. 1994. Visual perception and visual awareness after brain damage: A tutorial overview. In (C. Umilta and M. Moscovitch, eds.) *Consciousness and Unconscious Information Processing: Attention and Performance 15.* Cambridge: MIT Press.

Flohr, H. 1995. Sensations and brain processes. *Behavioral Brain Research* 71: 157-61.

Gray, J.A. 1995. The contents of consciousness: A neuropsychological conjecture. *Behavioral and Brain Sciences* 18: 659-722.

Greenfield, S. 1995. *Journey to the Centers of the Mind*. New York: W.H. Freeman.

Hameroff, S.R. 1994. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1: 91-118.

Leopold, D.A. & Logothetis, N.K. 1996. Activity-changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* 379: 549-553.

Libet, B. 1993. The neural time factor in conscious and unconscious events. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). New York: Wiley.

Llinas, R.R., Ribary, U., Joliot, M. & Wang, X.-J. 1994. Content and context in temporal thalamocortical binding. In (G. Buzsaki, R.R. Llinas, & W. Singer, eds.) *Temporal Coding in the Brain*. Berlin: Springer Verlag.

Logothetis, N. & Schall, J. 1989. Neuronal correlates of subjective visual perception. *Science* 245: 761-63.

Milner, A.D., and Goodale, M. A. 1995. *The Visual Brain in Action*. Oxford: Oxford University Press.

Shallice, T. 1988. Information-processing models of consciousness: possibilities and problems. In (A. Marcel and E. Bisiach, eds*.) Consciousness in Contemporary Science*. Oxford: Oxford University Press.

Sheinberg, D. L. and Logothetis, N.K. 1997. The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences USA* 94: 139-141.

Taylor, J.G. & Alavi, F.N. 1993. Mathematical analysis of a competitive network for attention. In (J.G. Taylor, ed.) *Mathematical Approaches to Neural Networks*. New York: Elsevier.

Tootell, R.B., Reppas, J.B., Dale, A.M., Look, R.B., Sereno, M.I., Malach, R., Brady, J. & Rosen, B.R. 1995. Visual motion aftereffect in human cortical area MT revealed by functional magnetic resonance imaging. *Nature* 375:139-41.