BEHAVIORAL AND BRAIN SCIENCES (1999) **22,** 809–869
Printed in the United States of America

A neuron doctrine in the philosophy of neuroscience

Ian Gold
Institute of Advanced Studies, Australian National University, Canberra ACT 0200, Australia
iangold@coombs.anu.edu.au
www.coombs.anu.edu.au/Depts/RSSS/People/IanGold.html;
Department of Ophthalmology, Royal Victoria Hospital, Room H414, 687 avenue des Pins ouest, Montreal, Quebec, Canada H3A 1A1
ian@vision.mcgill.ca

Daniel Stoljar
Department of Philosophy and Institute of Cognitive Science, University of Colorado, Boulder, CO 80309
stoljar@colorado.edu;
Institute of Advanced Studies, Australian National University, Canberra ACT 0200, Australia
dstoljar@coombs.anu.edu.au
www.coombs.anu.edu.au/Depts/RSSS/People/Stoljar.html

Abstract: Many neuroscientists and philosophers endorse a view about the explanatory reach of neuroscience (which we will call the *neuron doctrine)* to the effect that the framework for understanding the mind will be developed by neuroscience; or, as we will put it, that a successful theory of the mind will be solely neuroscientific. It is a consequence of this view that the sciences of the mind that cannot be expressed by means of neuroscientific concepts alone count as indirect sciences that will be discarded as neuroscience matures. This consequence is what makes the doctrine substantive, indeed, radical. We ask, first, what the neuron doctrine means and, second, whether it is true. In answer to the first question, we distinguish two versions of the doctrine. One version, the *trivial* neuron doctrine, turns out to be uncontroversial but unsubstantive because it fails to have the consequence that the nonneuroscientific sciences of the mind will eventually be discarded. A second version, the *radical* neuron doctrine, *does* have this consequence, but, unlike the first doctrine, is highly controversial. We argue that the neuron doctrine appears to be both substantive and uncontroversial only as a result of a conflation of these two versions. We then consider whether the radical doctrine is true. We present and evaluate three arguments for it, based either on general scientific and philosophical considerations or on the details of neuroscience itself, arguing that all three fail. We conclude that the evidence fails to support the radical neuron doctrine.

Keywords: Churchlands; classical conditioning; cognitive neuroscience; Kandel; learning; materialism; mind; naturalism; neurobiology; neuron doctrine; neurophilosophy; philosophy of neuroscience; psychology; reduction; theoretical unification

1. Introduction

Among those who reflect on the nature of neuroscience, there is a view about its scope and limits which we will call, with a certain amount of historical license, *the neuron doctrine.*[1] Roughly, the neuron doctrine is the view that the framework within which the science of the mind will be developed is the framework provided by neuroscience; or, as we put it, that a successful theory of the mind will be a solely neuroscientific theory.[2] The idea is not, of course, that neuroscience will explain everything about the mind; perhaps there are aspects of the mind we will never explain. The idea is rather that, to the extent that we will achieve a scientific understanding of mental or psychological phenomena at all, neuroscience will be the science that achieves it. According to the neuron doctrine, in the race to explain the mind, smart money is on neuroscience.

There are at least three reasons for thinking that the neuron doctrine so described is important. First, to claim that it will one day explain the mind is obviously one of the more fundamental and ambitious claims that could be made about any science. If, as the neuron doctrine alleges, neuroscience will explain the mind, that would make it tremendously important to scientists and nonscientists alike. Any discussion of the foundations of neuroscience, therefore, must involve an assessment of the neuron doctrine.

IAN GOLD is a fellow in the Vision Research Unit at McGill University. In January 2000 he takes up a faculty position in the Department of Philosophy at Monash University in Melbourne.

DANIEL STOLJAR is Assistant Professor of Philosophy at the University of Colorado, at Boulder, and a Research Fellow at the Institute of Advanced Studies, Australian National University, Canberra. He is the author of a number of articles in philosophy of mind, metaphysics, and meta-ethics.

© 1999 Cambridge University Press 0140-525X/99 $12.50

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

Second, the neuron doctrine seems to be strongly supported by science and philosophy. This might be brought out in the following way.[3] Many scientists and philosophers adhere to the methodological view known as *naturalism.* According to naturalism, to the extent that we will be able to understand the world, it will be empirical science (and not, for instance, religion or philosophy) that provides that understanding. Similarly, many scientists and philosophers adhere to the metaphysical view sometimes known as *materialism.* Roughly, materialism holds that psychological events, states, and processes are nothing more than events, states, and processes of the brain.[4] Given these two views, and treating neuroscience by definition as the science of the brain, it seems inevitable that the neuron doctrine is true: if the mind is the brain, and if neuroscience is the science of the brain, then it is practically[5] a fact of logic that neuroscience is the science of the mind, and that it alone will explain what *can* be explained about the mind. Indeed, from this point of view, it is difficult to deny the neuron doctrine without sounding – as the philosopher Frank Jackson (1982) has put it in a related context – like someone who believes in fairies.

Finally, proponents of the neuron doctrine often suggest that their view apparently has important, and potentially devastating, consequences for our current practice of attempting to construct a scientific understanding of the mind, and perhaps even for intellectual domains further afield, such as the structure of scientific theories, the correct approach to the understanding of the social world, and the proper conception of morality, art, and the self.[6] This is because, on the face of it, neuroscience is not *the* science of the mind, as the neuron doctrine suggests, for the simple reason that it is not the *only* science of the mind. On the contrary, there are plenty of apparently non-neuroscientific disciplines, such as psychology, psychophysics, linguistics, ethology, and the like – sciences we will group together as the *psychological sciences.* If the neuron doctrine is true, what are we to make of them? For proponents of the neuron doctrine, the inevitable result of tracing out its consequences is that the psychological sciences must be relegated to a second-rate, or placeholder, status. Of course, one might fail, or simply refuse, to draw this consequence, but, from the point of view of the neuron doctrine, this could only constitute a failure of intellectual nerve and, anyway, does nothing to undermine the importance of the doctrine for the status of these fields. In short, the neuron doctrine seems to be a remarkable thesis, one with solid intuitive foundations but with a strikingly counterintuitive result.

This target article is a critical examination of the neuron doctrine and the philosophy of neuroscience on which it is based, with a particular focus on the consequence for the psychological sciences that it apparently entails. Our central claim is that the doctrine suffers from a fatal ambiguity. Interpreted one way, the neuron doctrine is highly plausible and does find strong support in science and philosophy. However, on this interpretation, it fails to have the revolutionary consequence for the psychological sciences suggested by its proponents. Interpreted another way, the neuron doctrine is extremely interesting and would have this consequence, but we argue that there is little evidence that, on this interpretation, the doctrine is true. The problem with the neuron doctrine, we will claim, is that there is no way for it to be made both plausible and interesting.

We begin our examination of the neuron doctrine by asking who holds it and by considering in some detail the revolutionary consequence of doing so. In section 2, we distinguish two versions of the doctrine, one trivial and one radical, and we argue that these versions have not been clearly distinguished in the literature on neuroscience. We then turn to three arguments for the radical version of the doctrine, each prompted by scientific claims or views in the philosophy of science. The first argument, which we call *the argument from naturalism and materialism* (sect. 3), is based on a small number of highly plausible claims that we take most neuroscientists and philosophers accept. The second argument, which we call *the argument from unification* (sect. 4), attempts to defend the neuron doctrine by appealing to considerations about the development of scientific theories. The final argument, which we call *the argument from exemplars* (sect. 5), looks to neuroscience itself for support of the radical version of the neuron doctrine. To evaluate this argument, we consider one case of neuroscientific theory at some length, namely, the theory of elementary learning in the marine snail *Aplysia* from Eric Kandel and his colleagues. These arguments do not exhaust all the ones that might be offered in defense of the neuron doctrine, but we have chosen to discuss them because they are all highly plausible, because they appeal to principles respected by neuroscientists and philosophers alike, and because they point up important conceptual features of neuroscience and its place among the sciences of the mind. We conclude, in section 6, with some remarks about the morals one might draw from our argument.

1.1. Who holds the neuron doctrine?

For the purposes of our discussion, we will take the chief proponents of the neuron doctrine to be the philosopher-neuroscientists Patricia S. and Paul M. Churchland. There are two reasons for this. First, as we shall see, the Churchlands are particularly clear and knowledgeable advocates of the doctrine.[7] Second, because neuroscientists themselves tend to be reticent about expressing metascientific commitments in anything other than popular or quasi-popular publications (some of which we canvass below), it has largely been left to the Churchlands to articulate in a technical way the status and commitments of neuroscience. Indeed, more than anyone else on the contemporary scene, the Churchlands are responsible for painting the portrait of neuroscience and for rightly drawing attention to its many successes. Their advocacy of the doctrine can therefore be reasonably taken as a reflection of a central, perhaps dominant, intellectual trend in the field as a whole.[8]

A clear statement of the neuron doctrine can be found at the beginning of Patricia Churchland and Terrence Sejnowski's book, *The computational brain:* "The working hypothesis underlying this book is that emergent properties are high-level effects that depend on lower-level phenomena in some systematic way. Turning the hypothesis around to its negative version, it is highly improbable that emergent properties cannot be explained by low-level properties" (1992, p. 2). In explaining these remarks, Churchland and Sejnowski say that those who deny their hypothesis are making a certain kind of prediction about science and that

> as the history of science shows all too clearly, predictions grounded in ignorance rather than knowledge often go awry. In advance of a much more highly developed neurobiology than currently exists, it is much too soon to be sure that *psychological phenomena cannot be explained in terms of neurobiological*

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

> *phenomena.* Although a given phenomenon such as protein folding or awareness of visual motion cannot *now* be explained, it might yield to explanation as time and science go on. . . . Searching for reductive explanations of emergent properties does not entail that we should expect the explanations to be simpleminded, or breezily cobbled up or straightforwardly readable off the data points; it means only that the betting man keeps going. (1992, p. 3; emphasis added)

Because those who deny Churchland and Sejnowski's hypothesis are asserting that "psychological phenomena cannot be explained in terms of neurobiological phenomena," their own hypothesis, evidently, is that psychological phenomena *can* be so explained, or, at any rate, that this is a view one should take as a working hypothesis – a view that a betting man would lay money on. According to Churchland and Sejnowski, then, smart money is on neuroscience, and this is what makes them proponents of the neuron doctrine. As they put it in a more specific discussion of learning and memory, "[I]n the last analysis, the heart of the problem is to explain *global* changes in a brain's output, on the basis of orderly *local* changes in individual cells. That is, we want to discover how neuronal plasticity – a local property – can result in learning – a global property" (1992, p. 239).

A statement of the neuron doctrine can also be found at the beginning of Paul Churchland's (1995) recent book, *The engine of reason, the seat of the soul:*

> [R]ecent research into neural networks, both in animals and in artificial models, has produced the beginnings of a real understanding of how the biological brain works – a real understanding, that is, of how you work, and everyone else like you. . . . [W]e are now in a position to explain how our vivid sensory experience arises in the sensory cortex of our brains: how the smell of baking bread, the sound of an oboe, the taste of a peach, and the color of a sunrise are all embodied in a vast chorus of neural activity. We now have the resources to explain how the motor cortex, the cerebellum, and the spinal cord conduct an orchestra of muscles to perform the cheetah's dash, the falcon's strike, or the ballerina's dying swan. More centrally, we can now understand how the infant brain slowly develops a framework of concepts with which to comprehend the world. And we can see how the matured brain deploys that framework almost instantaneously: to recognize similarities, to grasp analogies, and to anticipate both the immediate and the distant future. (1995, pp. 4–5)

Although he concedes elsewhere in the book that the claim that neuroscience is *already* in a position to understand a number of mental phenomena is hyperbole,[9] it seems clear from this passage and others that Churchland endorses the neuron doctrine, for even if we are not currently in possession of a neuroscientific explanation of the mind, he is evidently confident that we will be. In an earlier paper, for example, P. M. Churchland says that his approach represents "an unabashedly reductive strategy for the neuroscientific explanation of a variety of cognitive phenomena," and that "the mystery of how the brain *represents* the world, and how it performs *computations* on those representations" appears to admit of a simultaneous solution as the "mystery of the brain's microphysical organization" (1989b, pp. 78–79).

One can also see a commitment to the neuron doctrine in the joint work of the Churchlands. In a discussion of how the relation of psychology to neuroscience is likely to compare with significant historical cases of intertheoretic reduction – the reduction of Kepler's planetary laws to Newton's laws of motion, the reduction of temperature to mean molecular kinetic energy, and of light to electromagnetic waves – the Churchlands say that in the neuroscience-psychology case

> the presumption in favor of an eventual reduction (or elimination) is far stronger than it was in the historical cases just examined. For unlike the earlier cases of light or heat or heavenly motions, in general terms we already know how psychological phenomena arise: they arise from the evolutionary and ontogenetic articulation of matter, more specifically, from the articulation of biological organization. We therefore *expect* to understand the former in terms of the latter. The former is produced by the relevant articulation of the latter. (1994, p. 48)

To say that we should expect a reduction of psychology to neuroscience, and therefore that we should expect to understand psychological phenomena in neuroscientific terms, is to say that we expect that a successful theory of the mind will be a solely neuroscientific theory. In other words, it is to endorse the neuron doctrine.

Although our primary focus is the Churchlands, commitment to the neuron doctrine is by no means limited to them. In *A vision of the brain,* for example, Semir Zeki says:

> It is . . . fortunate that neurobiologists are not philosophers, for they might otherwise find themselves immersed, like the philosophers, in an endless and ultimately fruitless discussion of the meaning of words such as "unconscious," or "inference" or "knowledge" and "information" instead of trying to unravel important facts about the brain. They would, in brief, end up contributing as meagrely to an understanding of the brain and of the mind as philosophers have. This last point is not a trivial one for ultimately the problems that cortical neurobiologists will be concerned with are the very ones that have preoccupied the philosophers throughout the ages – problems of knowledge, experience, consciousness and the mind – all of them a consequence of the activities of the brain and ultimately only understandable when the brain itself is properly understood. *The path toward the millennial future lies more with neurobiologists and some philosophers acknowledge this. . . . It is only through a knowledge of neurobiology that philosophers of the future can hope to make any substantial contribution to understanding the mind.* (1993, p. 7; emphasis added)

In a similar vein, Solomon Snyder says:

> Of all the momentous revolutions of twentieth-century science, two hold particular promise for bringing the mystery of human consciousness into the realm of human understanding. One of these revolutions is the development of new groups of drugs that produce extraordinary effects upon the mind; the other is the explosion in our understanding – at the cellular and molecular levels – of just how the human brain works. (1996, p. 1)

Similarly, Gerald Edelman, arguably the only neuroscientist who has attempted to formulate a comprehensive neuroscientific theory of the mind, says that his aim is to

> construct a scientific theory of the mind based directly on the structure and workings of the brain. By "scientific" in this context, I mean a description based on the neuronal and phenotypic organization of an individual and formulated solely in terms of physical and chemical mechanisms giving rise to that organization. (1989, pp. 8–9)

Also, Francis Crick writes: "The scientific belief is that our minds – the behavior of our brains – can be explained by the interactions of nerve cells (and other cells) and the molecules associated with them" (1994, p. 7).

Nor is the neuron doctrine a particularly recent view. In a paper written in 1974, David Hubel says that "the object of neurobiology is to understand the nervous system. . . . [T]his amounts to asking what happens in our heads when

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

we think, act, perceive, learn, or dream" (1974, p. 243). Also, in a classic paper written in 1972, Horace Barlow formulated a view he baptized "a neuron doctrine for perception," according to which "a picture of how the brain works, and in particular how it processes and represents sensory information, can be built up from knowledge of the interactions of individual cells" (1972, p. 384). Although Barlow's aim in that paper is to defend the view that perception will be explained at the level of *single* neurons rather than ensembles of neurons, his neuron doctrine is nonetheless also an instance of ours because it entails that perception will be explained at the neural level or at a level reducible to the neural: "A higher-level language than that of neuronal firing might be required to describe and conceptualize such [sensory and motor exploratory] games, but its elements would have to be reducible to, or constructible from, the interactions of neurons" (1972, p. 391).

In more recent work, Barlow reaffirms his commitment to the neuron doctrine, though in a more qualified way:[10]

> Wonder and astonishment at what the brain does are fully justified, but the reductionist attempt to explain its actions by the organized activity of individual nerve cells is not thereby doomed to failure, and the conclusion this chapter tends toward is that, although it has far to go, this theory is actually making steady progress. (1995, p. 416)

There is substantial evidence, therefore, that the neuron doctrine is widely held by philosophers and neuroscientists. We turn now to its radical consequence.

1.2. A consequence of the neuron doctrine

According to some of its proponents, the importance of the neuron doctrine lies in what it implies about the various sciences of the mind. In *Neurophilosophy,* Patricia Churchland writes: "[D]iscoveries in neuroscience will undoubtedly change out of all recognition a host of orthodoxies beloved in philosophy. Barring a miracle (or a calcified stubbornness), it will in particular transfigure epistemology, as we discover what it *really* means for brains to learn, to theorize, to know, and to represent" (1986, p. 482; emphasis added).

In speaking of the "host of orthodoxies beloved in philosophy," it is clear from the context that Churchland means, among other things, the tendency to try to explain the mind by invoking the psychological sciences, and to view such sciences as irreducible to neurobiology.[11] Moreover, it is Churchland's use of the word "really" in the passage that indicates the strength of her claim. What the passage suggests is that it follows from the neuron doctrine that the psychological sciences, and the epistemological views they support, cannot *really* tell us the truth, or the whole truth, about what it is to learn, to theorize, to know, and to represent. However, this in turn suggests that the psychological sciences are in the position of offering only superficial, partial, or inaccurate characterizations of the mind, and that these sciences will be superseded when neuroscience develops its own explanations of the phenomena. It is for this reason that neuroscience will change current orthodoxies "out of all recognition." Psychological explanations may be necessary way stations on the road to a successful theory, but they will not give us real insight into the phenomena they seek to explain.

A similar suggestion is made by Hubel:

> As we learn more about the brain, the effects of that knowledge on other fields of inquiry will be profound. The branches of philosophy concerned with such subjects as the nature of the mind and of perception will, in a sense, be superseded, as will the parts of psychology that seek to obtain the answers by indirect means. (1974, p. 259)

Here the suggestion that the psychological sciences are way stations is explicit. If a theory of a mental phenomenon is not a neural theory or, at any rate, is not reducible to the neural, it will be discarded in the long run by neuroscience as "direct" explanations become available.[12]

It is important to see just what a strong consequence this is. If the neuron doctrine implies that any psychological theory of the mind is second-grade, or placeholder, science, we are faced with the problem of what to say about the many developed psychological theories we now have. Linguistics, to take one example, is among the most advanced sciences of any mental phenomenon. According to many linguists, the fact that every normal human being is linguistically competent is to be explained by our (largely unconscious and innate) knowledge of a system of rules and principles that assign semantic and syntactic interpretations to physical forms, whether those forms are heard, seen, or touched. Linguists and psychologists are in the process of describing this system of rules, its development in childhood, and its interaction with other cognitive systems. However, according to the neuron doctrine, there is something misguided about this enterprise, because it is far from clear that the basic concepts of linguistic theory as it is currently understood can be reduced to the basic concepts of neuroscience as *it* is currently understood; indeed, many linguists believe they cannot (Higginbotham 1990). So linguistics faces a dilemma: either it must be reformulated in neuroscientific terms, or else it must be judged a placeholder science. In Churchland's terminology, if the neuron doctrine is true, linguistics does not tell us what it really means to have knowledge of language; and in Hubel's terminology, linguistics is in the position of providing only an indirect account of the phenomena it seeks to explain. Given the relative maturity and complexity of linguistic theory, this is no trivial result.[13]

Strong as this consequence is, the Churchlands quite explicitly accept it and regard it, in fact, as a major selling point of their view. Paul Churchland (1989a, p. 109; see also Churchland 1995), for example, writes that the position he advocates holds out "the possibility of an alternative to, or potential reduction of, the familiar Chomskyan picture." Patricia Churchland is even more explicit. Churchland and Sejnowski write:

> [I]n linguistics it may be useful as a first pass to characterize a speaker's knowledge of semantics in terms of lists stored in memory. If, however, we want to take the further step and ask how speech production and comprehension are *really* done, given a more neurobiologically realistic construal of memory, then the semantics-as-list is a caricature that must be replaced by something closer to the truth. And having a more neurobiologically realistic characterization of semantic memory may well result in new ways of looking at old data, and in new hypotheses that would not have seemed at home in the old framework. (1992, p. 416; emphasis added)

The radical consequence of the neuron doctrine has therefore been noticed by its advocates, and it is one of the features of the doctrine that lends it an air of intellectual excitement.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

1.3. Evidence for the neuron doctrine

Of course, one would have to accept this consequence, revolutionary though it is, if the neuron doctrine were backed up by the best empirical results. If we currently had a mature neuroscience that could explain a wide range of mental phenomena, then we would have to admit that the interpretation of linguistics and the other psychological sciences was settled. Is it the case then that the facts are in, but nobody has bothered to tell the linguists and psychologists?

Of course the answer is "no." Although we have a great deal of knowledge about the basic biology of the brain, it is only a slight exaggeration to say that we are almost completely ignorant about how the brain produces mental life. As we remarked above, Paul Churchland's optimism about current neuroscience should really be understood as hyperbole, and other neuroscientists are far more cautious. Hubel, for example, says that

> [t]he knowledge we have now is really only the beginning of an effort to understand the physiological basis of perception, a story whose next stages are just coming into view; we can see major mountain ranges in the middle distance, but the end is nowhere in sight. . . . We are far from understanding the perception of objects, even such comparatively simple ones as a circle, a triangle, or the letter A – indeed, we are far from even being able to come up with plausible hypotheses. (1988, pp. 219–20)

And what goes for the physiology of vision, where we have considerable understanding, also goes for the physiology of language and of our other cognitive capacities.

So we have a problem. On the one hand, the neuron doctrine has widespread support, in part because it seems to follow from widespread views in both philosophy and science. On the other hand, the neuron doctrine has a consequence concerning the effect of neuroscience on the psychological sciences that is not only radical but unsupported by neuroscience itself. The problem is how to explain the existence of a single doctrine that is simultaneously radical – and radical in the absence of evidence – as well as widespread.

1.4. Two questions

Our first question, therefore, is this: Given that the facts required to evaluate the neuron doctrine are not in, why do informed people believe it? Our answer to this question (set out in sect. 2) is that the neuron doctrine suffers from an ambiguity. Interpreted one way, the doctrine is very plausible but fails to have the radical consequence for the psychological sciences that we have discussed. Interpreted another way, the neuron doctrine is an empirical conjecture that *does* have this substantive consequence but is highly controversial and currently unsupported by the evidence. Our suggestion is that proponents of the neuron doctrine sometimes conflate these two very different interpretations of their view and believe as a result that there is a *single* view that is both obvious and revolutionary.

Our second question is this: Is there any scientific reason to believe the radical and interesting form of the neuron doctrine? We address this question in sections 3–5 by considering three arguments that might be offered in its defense.

2. Two versions of the neuron doctrine

We have expressed the neuron doctrine as the view that a successful theory of the mind will be a solely neuroscientific theory. What exactly does that mean? In the first part of this section, we argue that attention to this question leads to a distinction between two conceptions of neuroscience, which in turn makes it possible to distinguish two versions of the neuron doctrine. We then provide evidence that proponents of the doctrine frequently conflate these two versions.

2.1. Two conceptions of neuroscience

According to one conception of neuroscience, perhaps the more traditional conception, neuroscience is to be understood as the science we will call *biological neuroscience,* the concern of which is the investigation of the structure and function of individual neurons, neuronal ensembles, and neuronal structures. For simplicity, we will stipulate that biological neuroscience includes only neurophysiology, neuroanatomy, and neurochemistry, and we will take it to be synonymous with *neurobiology.*

According to another conception, neuroscience is taken to be what is often called *cognitive neuroscience* (see Gazzaniga 1995; see also Kosslyn & Andersen 1992 and Kosslyn & Koenig 1995), and we will adopt that name here. Cognitive neuroscience is an interdisciplinary approach to the study of the mind, the concern of which is the integration of the biological and physical sciences – including in particular biological neuroscience – with the psychological sciences to provide an explanation of mental phenomena. Although biological neuroscience is interested in understanding the biology of the brain, cognitive neuroscience attempts to synthesize biology and psychology to understand the mind. Cognitive neuroscience therefore includes biological neuroscience as a proper part but is not exhausted by it.

2.2. Versions of the neuron doctrine

With these conceptions of neuroscience before us, we can now distinguish two versions of the neuron doctrine. The two doctrines are generated by replacing "neuroscience" in our general statement of the neuron doctrine above by "cognitive neuroscience" and "biological neuroscience." We will call these versions the *trivial neuron doctrine* and the *radical neuron doctrine,* respectively.

2.2.1. The trivial doctrine. The trivial neuron doctrine is the view that a successful theory of the mind will be a solely cognitive neuroscientific theory. According to this doctrine, to the extent that psychological phenomena will be explained at all, the science that will do so is cognitive neuroscience. Because cognitive neuroscience includes any concept from the psychological or biological sciences (including any of the branches of physical science that might be relevant to describing the brain), the theory of the mind will turn out to involve any one of a very large number of possible combinations of scientific concepts. For example, the future of research could see the psychological sciences providing the functional description of the phenomena to be explained and biological neuroscience pro-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

viding the mechanistic account of how function is implemented in the brain.[14]

The essential feature of this version of the neuron doctrine is that it does not have the radical consequence for linguistics and the other psychological sciences discussed above. In the first place, this version does not entail that linguistics and the other psychological sciences will be superseded because it is consistent with the trivial neuron doctrine that psychological science will be part of the successful explanation of the mind. In the second place, and more important, this view entails nothing about the concepts that will be used in a successful theory. The claim that the theory of the mind will be expressed in cognitive neuroscientific terms expresses nothing more, therefore, than an ecumenism in the development of the theory and an agnosticism about its content.

The trivial neuron doctrine is therefore a very weak doctrine indeed. The picture that emerges from it has three components. First, the trivial neuron doctrine holds that the mind is a biological phenomenon; in other words, the trivial doctrine adheres to the thesis of materialism, the thesis that mental phenomena are neural phenomena. Second, the doctrine insists that the understanding of this phenomenon will derive from science; that is, the trivial doctrine adheres to the thesis of naturalism. Finally, however, the doctrine also holds that this understanding may not be provided by means of biological concepts alone but that psychological concepts may be required as well. Indeed, the trivial doctrine in principle leaves open which concepts will feature in the successful theory of the mind. Because a joint commitment to materialism and naturalism is a scientific commonplace, and because it has no radical consequences, we have called this version of the neuron doctrine the trivial doctrine.[15]

As one might expect, the trivial neuron doctrine is widely held by cognitive scientists. James Higginbotham, for example, says that although many cognitive scientists follow Descartes in supposing that

> the activities of the mind are not reducible to more familiar physical operations or to simpler mental activities shared by animals with and without language . . . [a]nother aspect of Descartes's conception of the mind is generally rejected as a research assumption . . . namely his thesis that the mind was a separable substance and in particular not a physical thing. Contrary to Descartes, cognitive scientists who otherwise adopt his views consider that the study of the mind is the study of the brain and nervous system, conducted at some level of abstraction that we would like to clarify. (1990, p. 249)

According to Higginbotham, because cognitive scientists believe that the study of the mind is the study of the brain, they count as cognitive neuroscientists in the sense we have described and hence as adherents of the trivial neuron doctrine. Following Higginbotham, we can say that the trivial neuron doctrine is committed only to the idea that a successful theory of the mind will be a theory of the brain – after all, it will not be a theory of the foot, or the kidney, or of an immaterial mind! – but neither must it be a theory of the brain expressed solely in terms of neurons and their properties.

In saying that the trivial neuron doctrine is trivial, however, we do not mean to suggest that it is compatible with any approach whatever to the study of the mind. In general, any theory of the mind that denies either (a) that the mind is a biological phenomenon, or (b) that the study of the mind is a part of natural science, or (c) that at least psychology or neurobiology is in principle relevant for the explanation of the mind, is incompatible with the trivial neuron doctrine. An example of the first (and perhaps the second) kind of theory is the version of *dualism,* according to which the mind is an object wholly distinct from the brain and body. An example of the second kind of theory is the version of *social constructivism* according to which the mind is a social construct in principle isolated from natural science. An example of the third kind of theory would be a certain version of the *artificial intelligence program,* according to which both neurobiology and the details of psychology are in principle irrelevant to the construction of theories of mentality in the most abstract sense. All such views are clearly incompatible with the trivial neuron doctrine.[16]

2.2.2. The radical doctrine. Because the trivial neuron doctrine amounts only to the claim that a successful theory of the mind will be a theory of the brain, it is uncontroversial and deserves to be as widespread as the neuron doctrine is. However, this version of the neuron doctrine is uninteresting because no radical, or even moderately substantive, consequences follow from it.

Nevertheless, there is a reading of the doctrine that does make it interesting. By substituting "biological neuroscience" for "neuroscience" in our formulation of the neuron doctrine, we get a radical doctrine according to which a successful theory of the mind will be a solely biological neuroscientific theory. Because, as we have said, we stipulate that biological neuroscience includes only neurophysiology, neuroanatomy, and neurochemistry, the radical neuron doctrine holds that neurophysiology, neuroanatomy, and neurochemistry will by themselves eventually have the conceptual resources to understand the mind and, as a consequence, a successful theory of the mind will make no reference to anything like the concepts of linguistics or the psychological sciences as we currently understand them.[17]

According to the radical neuron doctrine, a successful theory of the mind will be a theory of the brain expressed *in terms of* the basic structural and functional properties of neurons, ensembles, or structures. As a result, the radical neuron doctrine is substantive in having as its essential feature the consequence that the intellectual project pursued by Higginbotham and many others is doomed from the beginning. After all, Higginbotham assumes – and assumes that most cognitive scientists assume – that the rules posited by linguists and psychologists cannot be reduced to neurobiological notions. If Higginbotham and others are right about this, and if the radical neuron doctrine is true, then the psychological sciences will yield in the fullness of time to better biological neuroscientific theories. To adopt again the phraseology of Patricia Churchland and David Hubel, these sciences produce only *indirect* theories of the phenomena they seek to explain and do not tell us what these phenomena are *really* like.

2.2.3. Evidence of commitment to the radical doctrine. Once we have the distinction between the radical and trivial neuron doctrines clearly before us, the crucial question is whether the proponents of the doctrine intend to defend the radical or only the trivial version. It seems clear that at least *some* of the passages cited above expressing commitment to the neuron doctrine should in fact be taken as ex-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

pressions of commitment to the radical version of the doctrine.

Recall, for example, that in their discussion of the relation between psychological phenomena and neuroscience, the Churchlands write: "We therefore expect to understand the former in terms of the latter" (1994, p. 48). Also, Paul Churchland expresses a commitment to "an unabashedly reductive strategy for the neuroscientific explanation of a variety of familiar cognitive phenomena" (1989c, p. 78).

Churchland and Sejnowski say that "it is highly improbable that emergent properties cannot be explained by low-level properties" (1992, p. 2).

Also, here is Crick again: "The scientific belief is that our minds – the behavior of our brains – can be explained by the interactions of nerve cells (and other cells) and the molecules associated with them" (1994, p. 7).[18]

Finally, Barlow says, "[A] picture of how the brain works, and in particular how it processes and represents sensory information, can be built up from knowledge of the interactions of individual cells" (1972, p. 384).

2.3. Conflating the trivial and the radical doctrines

Although support for the neuron doctrine – and, in some cases, the radical doctrine – appears to be widespread, in our view supporters of the doctrine do not always distinguish between the two versions we have identified. Indeed, a closer examination of some of the texts we have considered reveals a tendency to conflate the two versions of the doctrine.

From the passages cited above, for example, it seems clear that the Churchlands' official view is the radical neuron doctrine. Nevertheless, they sometimes present the radical doctrine as equivalent to the trivial doctrine. Consider again the passage from Churchland and Sejnowski cited above: "The working hypothesis underlying this book is that emergent properties are high-level effects that depend on lower-level phenomena in some systematic way. Turning the hypothesis around to its negative version, it is highly improbable that emergent properties cannot be explained by low-level properties" (1992, p. 2).

Given the distinction we have introduced, it is clear that, contrary to what Churchland and Sejnowski obviously intend, the two hypotheses mentioned in this passage are by no means equivalent: the second is not the negative version of the first. The second claim says that it is highly probable that emergent psychological properties can be explained by low-level neurobiological properties. This is the radical neuron doctrine. The first claim, in contrast, says only that emergent psychological properties *depend on* low-level neurobiological properties in some systematic way. However, to say this is to say something extremely weak. In particular, it is not to say anything that a proponent of the trivial neuron doctrine need deny. Churchland and Sejnowski therefore conflate in this passage the trivial and the radical neuron doctrines by taking them to be nothing more than two formulations of the same claim.

To take a different example, consider the passage from the Churchlands that we quoted earlier:

> [I]n general terms we already know how psychological phenomena arise: they arise from the evolutionary and ontogenetic articulation of matter, more specifically, from the articulation of biological organization. We therefore *expect* to understand the former in terms of the latter. The former is produced by the relevant articulation of the latter. (1994, p. 48)

The interpretive difficulty presented by this passage is that, on the face of it, the argument implicit in it is invalid. The premise of the argument is that psychological phenomena arise from the brain, or, as the Churchlands put it, psychological phenomena are produced from the articulation of biological organization. The conclusion of the argument is that one should "understand the former in terms of the latter." However, this conclusion does not follow, and for a reason that the Churchlands are certainly aware of (see, e.g., Churchland & Churchland 1996, pp. 219–22). The mere fact that *A*s are made up of *B*s does not entail that we should expect an understanding of the *A*s in terms of the *B*s. Earthquakes, for example, are constituted by a set of causal processes involving the myriad microphysical particles that make up a swath of terrain, but these processes do not figure in any sensible explanation of the large-scale event (cf. Putnam 1975, pp. 295–97). So why should one think that the fact that psychological phenomena are produced by the brain has much to do with the theory of those phenomena? We suggest that the Churchlands are here conflating the two versions of the neuron doctrine, and as a consequence are defending a radical doctrine for reasons that only support a trivial doctrine. The fact that psychological phenomena are produced from an articulation of biological organization gives us a clear reason to believe that the theory of psychological phenomena will be a theory of the brain, and perhaps even a prima facie reason to expect that the theory of that organization will play some role in a successful theory of those phenomena; in other words, this fact gives us a reason to adopt the trivial neuron doctrine. However, the fact that psychological phenomena are produced by the brain does *not* give us a reason to adopt the radical neuron doctrine – to suppose that psychological phenomena are to be understood solely in neural terms. One could therefore derive the conclusion from the premise in the present argument only by conflating the two versions of the neuron doctrine.

This conflation is also evident if one examines central trends in the Churchlands' work as a whole. One such trend is the methodological idea that neurobiology is, or should be, relevant to the task of explaining cognitive or psychological phenomena. Patricia Churchland (1986), for example, emphasizes a "coevolutionary strategy" in developing theories of mental function that would explicitly take results in neurobiology into account (see also Churchland & Sejnowski 1992, p. 11). Also Paul Churchland (1990) writes of the need for "empirical and theoretical research into *brain* function in order to answer the question of what are the most important forms of representation and computation within cognitive creatures" (p. 158). He goes on to say that "the long-standing disinterest in the neurosciences, both within AI and in cognitive psychology . . . has been most unfortunate" (p. 200). Furthermore, the Churchlands describe the central fact that divides them from their critics as their rejection of what they call the "autonomy" of psychology (see, e.g., Churchland & Churchland 1996, p. 220), according to which the development of psychology is conceptually isolated from results in neurobiology.

Now claims of this kind are unobjectionable because they entail only that in developing a functional theory of a particular psychological phenomenon, one does well to keep an eye on whether the theory is likely to have some neuro-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

biological instantiation. Functional theories, as it is sometimes put, ought to be neurobiologically realistic, but what follows from this? Only that some approaches in psychology and artificial intelligence are mistaken. Although the Churchlands may be quite right in criticizing these approaches, it does *not* follow from these claims – contrary to what Paul Churchland (1990) goes on to say – that "fundamental insights into the general nature of cognition are likely to be found by examining the microstructure and microactivity of biological brains" (p. 225). Nor does it follow, as Patricia Churchland writes, that

> *[t]he co-evolutionary development of neuroscience and psychology means that establishing points of reductive contact is more or less inevitable.* As long as psychology is willing to test and revise its theory and hypotheses when they conflict with confirmed neurofunctional and neurostructural hypotheses, and as long as the revisions are made with a view to achieving concord with a lower-level theory, then the capacities and processes described by psychological theory will finally find explanations in terms of neuroscientific theory. (1986, p. 374)[19]

In emphasizing the methodological relevance of neurobiology, the Churchlands' position supports the trivial neuron doctrine. However, methodological relevance is much weaker than the explanatory sufficiency demanded by the radical neuron doctrine. Methodology only appears to support the radical neuron doctrine when the radical and trivial doctrines are run together.[20]

One can find a similar pattern of argument in other writers as well. In a passage we cited above, Edelman says that his aim is to "construct a scientific theory of the mind based directly on the structure and workings of the brain. By 'scientific' in this context, I mean a description based on the neuronal and phenotypic organization of an individual and formulated solely in terms of physical and chemical mechanisms giving rise to that organization" (1989, pp. 8–9). The claim that a theory of the mind will be formulated solely in terms of the physical and chemical mechanisms of the brain giving rise to the neuronal and phenotypic properties of the individual is strongly reminiscent of the Churchlands' claim that the explanation of the mind will amount to an explanation of the biological articulation of matter. However, immediately after making this claim, Edelman explicates it by saying that the theory of the mind

> must rest on a number of other psychological and physiological models, each of which is intricate and subject to error at our current stage of knowledge: models of perceptual categorization, memory, learning, concept formation, and, finally, language. The usual reductionist simplifying criteria – Occam's razor or a minimal number of assumptions – cannot usefully be applied to any such multilevel global model which must take into account a large series of evolutionary developments. (1989, pp. 9–10)

Our distinction between the trivial and the radical neuron doctrines makes it clear that this gloss on the initial formulation is altogether different from the initial formulation itself. Although the first passage suggests that the mind will be understood solely in terms of the basic biology of the brain, the second passage claims that the theory of the mind will not in fact be restricted to neurobiology but will require for its formulation a wide range of concepts from the psychological sciences.

Finally, in a passage quoted above, Zeki says:

> [U]ltimately the problems that cortical neurobiologists will be concerned with are the very ones that have preoccupied the philosophers throughout the ages – problems of knowledge, experience, consciousness and the mind – all of them a consequence of the activities of the brain and ultimately only understandable when the brain itself is properly understood. *The path toward the millennial future lies more with neurobiologists and some philosophers acknowledge this. . . . It is only through a knowledge of neurobiology that philosophers of the future can hope to make any substantial contribution to understanding the mind.* (1993, p. 7; emphasis added)

Zeki is arguing that features of the mind are the result of the activity of the brain and that the problems of the mind will only be solved when the brain is properly understood. These claims of course are uncontroversial; they express the trivial neuron doctrine. On the basis of these claims, however, Zeki (1993) asserts that "it is *only* through a knowledge of neurobiology that philosophers of the future can hope to make any substantial contribution to understanding the mind." But, as we have seen, this does not follow at all. One could argue that the relevance of the brain to the theory of the mind entails that neurobiology is the only way to understand the mind only if one fails to distinguish the trivial and radical doctrines.

2.4. The importance of the ambiguity

The fact that the neuron doctrine is ambiguous between at least the two claims we have identified is enormously important for understanding and evaluating the doctrine. What the ambiguity explains is why a view that is apparently radical and controversial is so widespread. The neuron doctrine is both widespread and controversial because it has one interpretation that renders it very plausible but unsubstantive, and one interpretation that renders it radical but unsupported by the scientific evidence. Our first question was why informed people believe the neuron doctrine in the face of inadequate evidence. Our answer to this question is that the doctrine is ambiguous, and running together two different versions of the doctrine gives it the illusory appearance of having the important features of both.

However, there is also another reason why the ambiguity is important. When confronted with the ambiguity, proponents of the neuron doctrine face two options: they can either say that they endorse only the trivial version of the doctrine, or they can stick their necks out and endorse the radical version. For the Churchlands, however, the first of these options is out of the question because the trivial doctrine has none of the consequences that the Churchlands clearly want to defend. As we have seen, the trivial neuron doctrine is no more than scientific common sense. Indeed, if the Churchlands, or anyone else, did intend to defend the trivial doctrine, that intention would be entirely mysterious. Why would one bother to defend explicitly a doctrine that is trivial or, at any rate, that just about everyone in the field accepts? To say that smart money is on neuroscience, in our trivial sense of the claim, is no more than to bet that *some* science of the mind or brain will win the race to understand the mind. However, that is not a bet that any rational bookie would take.

On pain of triviality, then, proponents of the neuron doctrine must adopt the radical version of the doctrine. However, this raises the second of the two questions we asked earlier, that is, whether the radical – and interesting – version of the doctrine is true. Our goal in the remainder of this target article is to try to answer this question by consider-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

ing three arguments for this version of the doctrine. In each case, we will suggest that the argument does not support the radical neuron doctrine. If we are right, the choice between the two versions of the neuron doctrine constitutes a destructive dilemma for its proponents: either to hold a view for which no scientific defense has been given, or to defend a view that requires no defense.

3. The argument from naturalism and materialism

The first argument we will consider is one that we have already mentioned in passing a number of times and that we will call the *argument from naturalism and materialism.* The attraction of this argument is that its premises are widely accepted both by neuroscientists and philosophers.[21] The first premise of the argument expresses commitment to what we have identified as *naturalism.* For our purposes, naturalism can be taken to be expressed by the following claim:

1. A successful theory of any class of natural phenomena is, or will be, provided solely by a developed science of those phenomena.

Naturalism therefore rules out the use of nonscientific methods of investigation in domains that have a science. The second premise of the argument expresses (a version of) the thesis of *materialism,* which we take to be supported by current science:

2. Mental phenomena are identical to neural phenomena.

Roughly, what (2) asserts is that when Kramer is excitedly anticipating the arrival of Mackinaw peaches, for example, his excitement is to be accounted for by the fact that something is going on in his brain or that his brain is in a particular state, and similarly for our claims that Kramer appreciates old movie theatres, that he is interested in coffee tables, that he prefers briefs to boxers, and so on. The final premise of the argument is a definition:

3. The science of neural phenomena (i.e., the science of the brain) is neuroscience.

From these premises one might reasonably infer:

4. A successful theory of mental phenomena is, or will be, provided solely by a developed neuroscience.

And (4) can be rewritten to express the neuron doctrine:

4′. A successful theory of mental phenomena will be a solely neuroscientific theory.

We have, then, an argument for the neuron doctrine that is based on one very appealing methodological position, one obvious empirical truth, and one innocent definition. How could one object to an argument like this?

3.1. The limitations of the argument

One objection that could be brought against it concerns the inference to premise 4. We have already seen that the fact that *A*s are made up of *B*s does not entail that the explanation of the *A*s is to be given in terms of the *B*s. In particular, the fact that mental phenomena are identical to neural phenomena does not entail that the *science* of mental phenomena is the *science* of neural phenomena.

There is also a second objection to this argument, one that concerns the distinction between the two conceptions of neuroscience that we have introduced. Even if the argument is successful, it is not an argument for the view of interest because premise 4′ is the ambiguous formulation of the neuron doctrine rather than the radical formulation. The conclusion we need is:

4*. A successful theory of mental phenomena will be a solely biological neuroscientific (i.e., neurobiological) theory.

However, 4* is a very different doctrine from 4′; 4* is, whereas 4′ is not, a substantial claim about the course of future science. In particular, as we have seen, 4* entails that the cognitive scientists described by Higginbotham are on entirely the wrong track. The view expressed by 4′, in contrast, entails no such thing. Most cognitive scientists believe they are theorizing about the brain, and are, according to premise 3, neuroscientists in the cognitive sense. Even if we agree, therefore, that 1, 2, and 3 entail the trivial neuron doctrine, the argument from these premises to 4* – the radical neuron doctrine – is invalid.

In response, one might claim that our definition of the science of the brain is incorrect. It is often assumed that the brain is studied by neurobiology so that the third premise actually amounts to this:

3*. The science of neural phenomena (i.e., the science of the brain) is biological neuroscience.

Moreover, from premises 1, 2, and 3* one *might* reasonably infer 4*. However, on this interpretation, the original argument is no longer uncontroversial because the third premise, 3*, is no longer innocent. It now begs the question against scientists and philosophers who take themselves to be studying the brain even though they are not studying the neurobiological properties of the brain. The scientists and philosophers in this class would reject 3* and the argument from 3* to the radical neuron doctrine.[22]

Once we have the ambiguity of the neuron doctrine clearly before us, therefore, a plausible argument for the doctrine turns out to be only an argument for its trivial incarnation. As we claimed in section 1, we suspect that many scientists and philosophers accept the neuron doctrine because of their commitment to some form of naturalism and materialism. Given that the trivial neuron doctrine is, as we have suggested, essentially an expression of those commitments, this is not surprising. However, although it can seem irresistible from these premises to draw the conclusion that the science of the mind will be solely neurobiological, with the two versions of the neuron doctrine in mind, it becomes clear why this line of argument is unpersuasive.

3.2. The trivial doctrine, naturalism, and materialism

We should emphasize that in our discussion of the argument above, we have not taken issue with the doctrines of naturalism and materialism. For the purposes of this article, we agree with the naturalism of the first premise, and we agree also that mental phenomena are identical to neural phenomena. In addition, we take it as obviously true that the mind is to be explained by appealing to the structure and function of the grapefruit-sized things in our skulls.[23] Our objection is only to the view that the best description of that thing will be entirely neurobiological and to the idea that naturalism and materialism provide support for such a view.

Indeed, if there is any position that sits uneasily with naturalism, it is not ours but the position of those who support the radical neuron doctrine. The only way to infer the radical doctrine from naturalism is to make a prediction about

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

the explanatory force of neurobiology in the fullness of time – a prediction we presented in the form of premise 3*. However, naturalism itself cannot support such a prediction. On the contrary, in deferring to science for judgments about what there is and what it is like, the naturalist ought to avoid predictions of this sort.

4. The argument from unification

We turn next to an argument for the neuron doctrine that appeals to the notion of unity in science as a marker of successful theories.

It is often suggested that the fact that a science can be unified or integrated with other sciences is a virtue, and that, as a consequence, science itself is tending toward unity (Oppenheim & Putnam 1958; Sellars 1963; but see Dupré 1993 for an extended argument to the contrary).[24] One reason for this view is that science attempts to construct a picture of the world, and the more coherent, seamless, and simple that picture is, the better. Another reason is the apparent connection between unity and explanation. The history of science seems to reveal that sciences or theories that unify previously disparate domains tend simultaneously to provide highly successful explanations of those domains; in unification, one finds explanation. The unification of electricity and magnetism in special relativity, the unification of gravity and inertia in general relativity, and the unification of evolutionary theory and genetics in neo-Darwinism are familiar examples. (We leave aside the question of whether unification *itself* constitutes a form of explanation, or whether successful explanations tend to *co-occur* with unification.)

Let us suppose then that unity is an important general tendency of science. How does this occur? A natural suggestion is that global unity in science will be the product of many local theoretical unifications in the various branches of science: think globally, act locally! It is here that the argument from unification for the radical neuron doctrine begins to take shape. The general idea is, within the cluster of sciences that deal with mental phenomena, there is one science that is best placed to support the global tendency toward unity, and that is neurobiology. Because of its obvious connections to biology[25] and thereby to the rest of science, a neurobiological theory of the mind would contribute most to the overall goal of unity, and this means that our best bet is to regard neurobiology as the eventual science of the mind.

We can make this argument more precise as follows. Its first premise is simply the presumed historical fact about science:

1. Science tends toward unity.

The second premise of the argument asserts that, in the context of the sciences of the mind, the science that will contribute most to global unity is neurobiology:

2. In the sciences of the mind, this tendency of science would be maximally supported by a unification of neurobiology and the psychological sciences.

The support for this premise derives from the idea that only a neurobiological unification would exhibit the psychological sciences as part of biology and of the rest of natural science. From these two premises, we may derive the conclusion that:

3. Science is tending toward a neurobiological theory of the mind.

Or rather:

3′. A successful theory of the mind will be solely neurobiological.

Of course, 3′ is simply the radical neuron doctrine. In other words, the argument from unification takes us from very general principles about science to the conclusion that the radical neuron doctrine is true.

Now in this formulation, the argument from unification is obviously open to a number of different objections. First, the idea that unification is a general trend of science is, as we have indicated, controversial in some quarters (Dupré 1993). Second, it is certainly not obvious – contrary to what the argument assumes – that the best way of achieving global unity is by means of local unifications.

We have some sympathy with both of these objections, but we think there is another objection that cuts deeper than either of them. The central problem with the argument from unification is that it does not distinguish among different conceptions of unification. Following Maudlin (1996), we distinguish three different relations or processes that might be denoted by "unification."[26] This in turn suggests that the idea behind premise 2 has at least three different formulations. We will argue that for none of the formulations is the argument persuasive.

4.1. Unification as dissolution

On the first interpretation, "unification" denotes a process we shall call *dissolution.* When dissolution occurs, the distinction between two theoretical domains is dissolved by a conceptual advance. That advance reveals the two domains to be features or manifestations of a single theoretical domain or to be derivable from that domain. In his paper, Maudlin cites special and general relativity as two paradigms of dissolution in this sense.

In the sciences of the mind, dissolution is also possible. It would require the discovery or development of a family of concepts that would reveal the biological features of the brain and the psychological features of the mind to be manifestations of, or derivable from, some third set of things. This view has traditionally been called *double-aspect theory* in philosophy, and it is surprising that it has so few contemporary supporters. As Maudlin notes, the likelihood of a unification of the four forces of nature in a "theory of everything" is a dogma in physics, but the parallel view in the philosophy of mind is thought to be eccentric.

One philosophical view that might be classified as a dissolution proposal is Wilfrid Sellars's (1963; 1971) theory of color. Sellars argued that the commonsense picture of the world (the *manifest image)* must be mistaken in its commitment to the existence of color because science has revealed that the world is made of atoms, and atoms cannot be colored. Sellars eliminated color as it is normally conceived from his manifest ontology, but he argued that in the completed scientific description of the world (the *scientific image)*, colors would be reintroduced in the form of entities he called *pure processes.* Sellars thought that pure processes would be discovered by physics, and for this reason color would be scientifically explicable. Being pure processes, however, colors would not have to be properties of colorless atoms. Pure processes would thus dissolve the distinction between the psychology of color and color physics, producing a third unified domain.

Sellars's account appeals to a future physics for a dissolu-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

tion of the boundary between the mental and the physical, but one might be inclined to think it more likely that neurobiology itself will produce radically new concepts that will dissolve that boundary. Indeed, this is the claim behind premise 2 of the argument from unification, if by "unification" one means dissolution. If the premise is interpreted this way, and the argument is persuasive, then the final theory of the mind will be neurobiological, and the radical neuron doctrine would be vindicated.

In our view, however, the argument is *not* persuasive when the premise is so interpreted, and for two reasons. First, suppose that future science discovers a radically new family of concepts that dissolves the distinction between neurobiology and psychology. It seems completely arbitrary to count these concepts as part of neurobiology rather than psychology; after all, the family of concepts is radically different from both. Moreover, once the concepts are in place, there will be no difference between these disciplines. Thus, dissolution undermines the radical neuron doctrine by doing away with neurobiology at the same time as it does away with psychology.

There is also a simpler reason why this version of the unification argument is unpersuasive. As Maudlin (1996) argues, dissolution is very hard to come by even in physics.[27] The theory of everything, like the radical neuron doctrine, is a dogma without much scientific support. Because physical theories rarely achieve dissolution, and because there is doubt whether fundamental physics will ever do so, even though a deep understanding of the phenomena is already available there, we conclude that there is that much less reason to think that the distinction between the neural and the mental will be dissolved by future theoretical advances. Because we doubt that dissolution will ever be achieved at all, we doubt a fortiori that neurobiology will produce it.

4.2. Unification as reduction

The second and most obvious possibility is that by "unification" one means *intertheoretic reduction.* Reductionism is the view that the concepts and the laws of a more basic theory – the *reducing theory* – can be used to capture and explain the phenomena described in a less basic theory – the *reduced theory.* In cases of reduction, the reduced theory is derived from, and exhibited as a proper part of, the reducing theory (for the locus classicus, see Nagel 1961; and for a more recent discussion, see the papers in Kim 1993).[28] In the case of psychology and neurobiology, this means that a science recognizable as neurobiology will eventually produce concepts that reduce psychology, just as physics reduced the concept of temperature to the concept of mean kinetic energy. Because neurobiology is taken to be more fundamental, the explanatory power of the theory would lie with neurobiology and would justify the claim that the successful theory of the mind was solely neurobiological.

A variant on this position is *eliminativism,* usually associated with the Churchlands (see, for example, P. M. Churchland 1981). Eliminativism envisages a replacement of psychology by neurobiology rather than a process by which psychology is exhibited as a proper part of neurobiology. Physics did not reduce the mistaken concept of phlogiston to some more fundamental physical concept; it disposed of the concept altogether. Similarly, on the eliminativist picture, a future theory of the mind will dispose of psychological or mental concepts. Although reduction and elimination appear at first blush to be rather different views, they are in fact two ends of a continuum defined by the extent to which one believes that psychology is correct in its description of the mind. At one end of the continuum is the view that psychology is entirely correct but that its description is not given in fundamental terms. In this case, psychology must be reduced to the level of neurobiology. At the other end of the continuum is the view that psychology is entirely mistaken about the mind and must be replaced by a neurobiology that starts from scratch with a new set of concepts. A more likely outcome of scientific progress is that some of psychology will be reduced and some eliminated as neurobiology develops. For our purposes, the class of views on the reduction-eliminativism continuum can be considered together. For simplicity, we will call this class of views "reduction."

Reduction differs from dissolution in the degree of radicalness envisaged in the future of neurobiology. Dissolution imagines that neurobiology will undergo an Einsteinian or Sellarsian revolution and obliterate the distinction between what we currently take to be neurobiology and psychology, whereas reduction envisages a less radical option in which neurobiology goes on much as it is but gradually develops the resources to flesh out or replace psychology. On the reductionist view, future neurobiology will be recognizable as neurobiology but will have greater explanatory power.[29]

There is an enormous literature dealing with reductionism, and it is not our concern to evaluate all the arguments for and against it here.[30] Rather, we are interested only in asking whether the argument from unification is persuasive if in premise 2 one reads "reduction" for "unification." As we have noted, the support for the premise is that considerations of unification privilege neurobiology over other sciences when it comes to explaining psychological phenomena. If by "unification" one means "reduction," this becomes the idea that considerations of reduction privilege biological neuroscience when it comes to explaining psychological phenomena. But do they?

In our view, the answer to this question is "no." If reductionism is a constraint in science at all, then it is a *general* or *global* constraint. That is to say, if reductionism is true, then anything that is in principle reducible, reduces to the most basic science there is, namely, physics.[31] However, this means that considerations of reduction do not privilege neurobiology over the other sciences but only physics, and reductionism implies that a successful theory of the mind will be solely physical and not solely neurobiological. The relation between neurobiology and psychology is left entirely open by reductionism and must be regarded as an empirical question about the local relations among the sciences. In short, then, the appeal to reductionism does too much for the proponent of the neuron doctrine. If one is going to be a reductionist, one has to take the train of reduction to the terminus of physics. In the absence of a further argument privileging neurobiology *as well as* physics, neurobiology represents for psychology nothing more than – in Fodor's (1981) phrase – a local stop. If there *is* an argument privileging neurobiology in this way, it is not an argument that derives from reductionism itself, and this is enough to defeat the argument from unification on the interpretation we are considering.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

4.3. Unification as conjunction

The final possibility we will consider is that by unification one means what we shall call *conjunction.* Conjunction is the process whereby two theories are unified simply by being joined together into a single larger theory. Although conjunction represents an extremely weak version of unification, it is not empty. For one thing, to conjoin two theories, they must at least be mutually consistent, and their consistency may not be a simple matter to establish. For another, the notion of conjunction must somehow be made interesting enough so that, as Maudlin (1996) puts it, merely showing that a theory of embryonic development and a theory of the formation of the rings of Saturn are not inconsistent when conjoined does not count as unifying them. However these issues are resolved, the important issue for us is that the claim of conjunction with respect to the mind is that a successful theory of the mind will be neurobiology-conjoined-with-psychology.

It is not necessary for us to decide here whether current neurobiology and psychology are consistent or, if not, whether they will one day be made consistent. It suffices for us to note that if, in premise 2, one means "conjunction" by "unification," then the argument from unification will not support the radical neuron doctrine because, unlike dissolution and reduction, conjunction unifies without doing away with anything, including psychology. Because psychology would continue to be part of a successful theory of the mind on this view, conjunction supports only the trivial, but not the radical, neuron doctrine.

5. The argument from exemplars

The two arguments for the radical neuron doctrine that we have considered thus far proceed from philosophical considerations. The final argument we discuss makes use instead of the details of neuroscience itself. We call this argument the *argument from exemplars.*

5.1. An inductive strategy

It might seem that any argument of this kind is doomed from the start because, as we noted at the beginning of this article, neuroscience is at an early stage of development. How then can an argument based on an embryonic neuroscience be developed?

The following passage from Hubel suggests a way in which this might be done:

> The brain has many tasks to perform, even in vision, and millions of years of evolution have produced solutions of great ingenuity. With hard work we may come to understand any small subset of these, but it seems unlikely that we will be able to tackle them all. It would be just as unrealistic to suppose that we could ever understand the intricate workings of each of the millions of proteins floating around in our bodies. Philosophically, however, it is important to have at least a few examples – of neural circuits or proteins – that we do understand well: our ability to unravel even a few of the processes responsible for life – or for perception, thought, or emotions, – tells us that total understanding is in principle possible, that we do not need to appeal to mystical life forces – or to the mind. (1988, p. 222)

Hubel's view leads to the suggestion that we may be able to infer the radical neuron doctrine by adopting the following inductive strategy. Let us suppose that there are pieces of neuroscientific theory that are each relatively successful at explaining a mental phenomenon. We will call these pieces of theory *exemplars.* If we take exemplars to be cases that are indicative of what a future theory of the mind will look like, we can construct an argument – or, rather, an argument schema – of the following form:

1. A successful theory of the mind will be made up of explanations of mental phenomena that are similar to, or have the same character as, exemplars, that is, current neuroscientific explanations of mental phenomena.

2a. Exemplar, e_1, provides an explanation of type *T* of a mental phenomenon.

2b. Exemplar, e_2, provides an explanation of type *T* of a mental phenomenon.

2c. Exemplar, e_3, provides an explanation of type *T* of a mental phenomenon 2n. Exemplar, e_n, provides an explanation of type *T* of a mental phenomenon.

Therefore,

3. A successful theory of the mind will be made up of explanations of type *T.*

If one wanted to devise an inductive argument to support the radical neuron doctrine, one would be able to do so if there are exemplars that are solely neurobiological – neuroscientific theories that explain mental phenomena solely by appealing to the concepts of neurobiology. By plugging "a solely neurobiological explanation" into the argument schema above, one produces a version of the argument from exemplars that supports the radical neuron doctrine:

1*. A successful theory of the mind will be made up of explanations of mental phenomena that are similar to, or have the same character as, exemplars.

2*a. Exemplar, e_1, provides a *solely neurobiological* explanation of a mental phenomenon.

2*b. Exemplar, e_2, provides a *solely neurobiological* explanation of a mental phenomenon.

2*c. Exemplar, e_3, provides a *solely neurobiological* explanation of a mental phenomenon 2*n. Exemplar, e_n, provides a *solely neurobiological* explanation of a mental phenomenon.

Therefore,

3*. A successful theory of the mind will be made up of solely neurobiological explanations.

That is,

3′. A successful theory of the mind will be a solely neurobiological theory.

If one thus assumes that a successful piece of neuroscientific theorizing gives us a window onto the future theory of the mind, and the view through that window reveals the theory to be solely neurobiological, then one can hold the radical neuron doctrine in advance of some of the evidence needed to support it. The task for a defender of this argument, therefore, is to find the right sort of exemplars – that is, neuroscientific theories that rely solely on neurobiological concepts.

5.2. Kandel's theory of learning in Aplysia

In order to evaluate the argument from exemplars, one would have to consider a number of exemplars in some detail.[32] In this article we will consider one – the neuroscientific theory of elementary learning developed by Eric Kandel and colleagues (Bailey & Kandel 1995; Castellucci & Kandel 1976; Hawkins et al. 1983; 1992; Hawkins & Kandel 1984; Kandel & Schwartz 1982; Pinsker et al. 1970; Wal-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

ters & Byrne 1983; Walters et al. 1981; for a summary see Kandel et al. 1995 and Shepherd 1994).

Although we consider only one exemplar, we take the upshot of the analysis of this case to be highly significant for the following reasons. First and foremost, this case appears to offer an account of some forms of naturally occurring learning in terms of neurons and their properties alone, and for this reason seems ideally suited to support the radical neuron doctrine. Moreover, the theory we consider has a number of other virtues that make it a worthy case to consider. It is a very successful bit of neuroscientific theory in the sense that it appears to offer a relatively complete account of a set of behaviors. It is successful also in the sense that it manages to integrate the behaviors in question not only with cellular neurobiology but also with the biochemical and molecular events that are crucial to those cellular processes. It thus traces learning to its most basic biology. Furthermore, Kandel's theory makes use of a central neuroscientific notion – the notion of *neural plasticity,*[33] according to which the structure and function of adult neurons can be altered – that plays an important role in learning theory quite generally and is related to other areas of neuroscience such as neural development. It is the sort of account, therefore, that might be expected to generalize at least to other forms of learning and possibly beyond.[34] Finally, we take it to be a sociological fact that Kandel's theory is widely regarded in the neuroscientific community as the best that neuroscience can now offer in the way of explanation of behavior or the mind in fundamental neuroscientific terms. If our critique of the radical neuron doctrine makes sense in the context of this bit of neuroscience, therefore, that is strong evidence for the account. Other exemplars would have to be examined to make the case complete, but this is a good place to start.[35]

5.2.1. Simple and associative learning. Kandel's theory deals with a cluster of elementary forms of learning called *simple learning* and *associative learning.* The theory attempts to explain these behaviors by appealing to the properties of a small family of neural circuits composed primarily of a sensory neuron-interneuron-motor neuron pathway, and the process of learning is hypothesized to be identical to a change in strength of the sensory neuron-motor neuron synapse. This change in synaptic strength occurs as a result of an alteration in the production of neurotransmitter in the sensory neuron. This basic model can be adapted or modified to account for a number of forms of learning including *habituation, sensitization,* and *classical conditioning,* and the features of the model may be able to account for other aspects of elementary learning as well (Hawkins & Kandel 1984).

Habituation is the process whereby a neutral stimulus is gradually ignored by an animal when it leads neither to harmful nor rewarding consequences. Sensitization is the process whereby a harmless stimulus that produces no aversive response comes to be experienced as noxious after being paired with a stimulus that naturally produces aversion. Both habituation and sensitization are classified as forms of simple learning in contrast to associative learning, the paradigm of which is classical conditioning – the learned association between two stimuli such as the ringing of a bell and the presence of food, in the well-known Pavlovian case. In classical conditioning, the spontaneous response to a nonneutral stimulus (the *unconditioned stimulus* or US) is transferred to the neutral stimulus (the *conditioned stimulus* or CS) in virtue of their contiguity, or repeated pairing in time.[36]

5.2.2. The neurophysiology of simple and associative learning. Because the neurophysiology of processes even as simple as these elementary forms of learning would be enormously difficult to study in complex organisms, Kandel and his colleagues use the marine snail *Aplysia californica* as a model. The *Aplysia* is an organism with a very simple central nervous system made up of about 20,000 cells, but it exhibits an innate gill-withdrawal reflex that can be modified in simple or associative learning paradigms.

For example, a neutral tactile stimulus to the tail of an *Aplysia,* initially producing a weak gill-withdrawal reflex, can be habituated. Habituation of this reflex occurs as a result of a decrease in the quantity of neurotransmitter released at the synapses made by sensory neurons on inter- and motor neurons. Sensitization in *Aplysia* occurs when a noxious stimulus causes a weak gill-withdrawal reflex to be strengthened. This process involves a change at the same neural locus as habituation, but in sensitization there is an *enhancement* of neurotransmitter release by the sensory neurons on their target cells. The process by which this occurs, however, is more complex. A mild stimulus to the siphon of an *Aplysia* produces a weak gill-withdrawal reflex. A shock to the tail activates facilitator interneurons that synapse near the synapse formed by the siphon sensory neurons on the motor neurons. The activity of the interneurons at this locus causes a molecular process to be initiated within the siphon sensory neuron, the upshot of which is that the sensory neuron is disposed to produce more neurotransmitter than before stimulation. When the siphon is weakly stimulated again, therefore, more neurotransmitter is produced, and a stronger gill-withdrawal reflex occurs. This process is called *presynaptic facilitation* because the sensory neuron (the *presynaptic neuron)* is made more effective, or *facilitated,* by means of the activation of the facilitator interneurons (Fig. 1).

The cellular mechanism of classical conditioning in *Aplysia,* is, on Kandel's model, an elaboration of the cellular mechanism of sensitization. In classical conditioning, a US, such as a tail-shock, is contiguous with a CS, such as a weak tactile stimulus to the siphon. Initially, only the US causes robust gill-withdrawal, but repeated pairings of the US and the CS eventually cause the gill-withdrawal to oc-

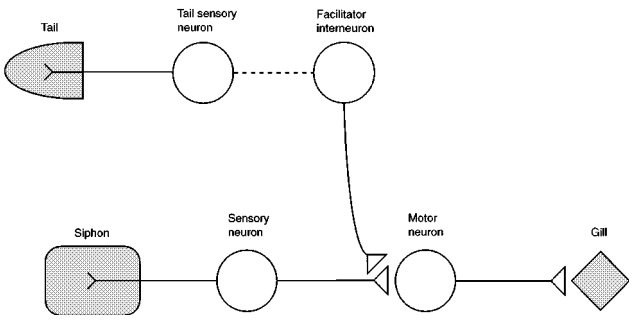


Figure 1. A partial circuit for sensitization and classical conditioning in *Aplysia.* In sensitization, a shock to the tail acts on siphon sensory neurons to bring about presynaptic facilitation. In classical conditioning, presynaptic facilitation is dependent upon the temporal pairing of conditioned stimulus (CS) (siphon) pathway activity and unconditioned stimulus (US) (tail) pathway activity. The CS pathway is active just before activation of the US pathway, thereby enhancing presynaptic facilitation of the siphon sensory neuron. (Adapted from Hawkins & Kandel 1984.)

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

cur in response to the CS as well. At the cellular level, the process of presynaptic facilitation in the pathway responding to the CS is enhanced by activity in the neural pathway responding to the US, which occurs just afterward. The action potentials generated by the US cause a greater facilitation of the sensory neurons responding to the CS as a result of the activation of the facilitator interneurons. This change makes the sensory neuron more effective in causing the motor neuron to fire and bring about gill-withdrawal in response to the CS. The primary difference between conditioning and sensitization, therefore, is the temporal coincidence of the activity of the facilitating and facilitated pathways. As a result, Kandel calls the process *activity-dependent presynaptic facilitation* (Fig. 1).

Before moving on to the philosophical questions at issue, it is important to observe that Kandel's model not only provides a neural description of the processes that underlie some forms of elementary learning, it also makes important conceptual contributions to the study of learning in general (see Kandel et al. 1995). Kandel argues that studies of learning in *Aplysia* reveal that short-term and long-term memory are part of a continuum and not two utterly distinct processes (Frost et al. 1985). The model also provides evidence that memory function in elementary learning is not a function of a neural network but of individual cells (Bailey & Kandel 1995).[37] Kandel's theory also holds that it is possible to produce a model of classical conditioning in terms of the process of sensitization, thus suggesting that the psychological distinction between simple and associative forms of learning may not be as hard and fast as one might suppose.[38] All of these are substantial claims about learning and not just its neurobiology.

5.3. Kandel's theory and the radical neuron doctrine

Clearly, then, Kandel's theory provides an admirable exemplar of neurobiological research. However, our aim here is not to evaluate the theory but to ask whether it supports the view that a successful theory of the mind will be solely neurobiological. It is important to be clear about what this question means. We are not asking whether Kandel's theory is true, let alone whether it constitutes a complete neuroscientific theory of learning, and still less of the mind! We are assuming that Kandel's theory is a successful explanation of the neurophysiology of elementary learning. What we want to know is whether, in accordance with the strategy of the argument from exemplars, Kandel's theory provides inductive support for the view that a future account of mental phenomena will be solely neurobiological. If we assume that Kandel's theory gives us a window onto a successful theory of the mind, can we infer inductively that the theory will be neurobiological?

Our answer to this question is "no." Whatever the virtues of Kandel's theory, we will argue that *it is not solely neurobiological,* and, for this reason, that the view through the window of that theory reveals the future theory of the mind not to be solely neurobiological either. The instance of the argument from exemplars that appeals to Kandel's theory, therefore, does not succeed. In what follows, we focus on the case of classical conditioning.

5.3.1. "Pure" neurobiology. We have claimed that Kandel's account is not purely neurobiological. In saying this, we mean that the notions involved in the description of classical conditioning in *Aplysia* include substantive psychological concepts and not merely the concepts of neurobiology. The history of classical conditioning, its roots in the associationism of Hume and Mill, and its incarnation in behaviorism is well known (Boring 1950; Pinker 1991).[39] Kandel's theory of classical conditioning is not developed in a vacuum but, as Kandel and coworkers acknowledge, makes use of recent work in psychology that is part of that tradition. This includes, in particular, the primary psychological model of classical conditioning – that of Rescorla and Wagner (1972; see also Gluck & Thompson 1987). In modeling *Aplysia* neurophysiology, that is, Kandel's theory appeals explicitly to psychological concepts,[40] and it is this fact that leads us to say that his model of conditioning is not solely neurobiological.

The fact that Kandel's account is developed within an explicit and highly theoretical psychological framework means that the account does not provide a genuine alternative to psychological theory. Rather, it absorbs the required psychological ideas in order to provide a framework for understanding the behavior of *Aplysia* neurons and their role in conditioning. Although Kandel's account changes our conception of conditioning somewhat (e.g., in providing evidence that it is constructed out of the mechanism of sensitization), it does not replace that theory. Its primary success is the fleshing out of a psychological story in neurobiological terms (Gluck & Thompson 1987). This should not be surprising. Because neurobiology has no concepts that can be used to describe the behavior of an animal, the notion of a "pure" neurobiology actively in competition with psychology can only be a vision of some future science.

In order to make our argument clearer, consider an illustration from the theory of color. One of the salient features of color phenomenology is the fact that only certain color combinations are possible. We can, for example, see reddish-blue and reddish-yellow but not reddish-green. This aspect of color perception is called *color opponency,* and it led Ewald Hering in 1877 to propose the *opponent process* theory of color perception, later revived and developed by Leo Hurvich and Dorothea Jameson (see Hurvich 1981). According to this theory, the space of all perceivable colors is organized along three axes of phenomenal difference – a red-green axis, a blue-yellow axis, and a black-white axis. The contemporary interpretation of opponent process theory appeals to opponent neural channels, the function of which depends on *color-opponent cells.* These neurons – different species of which exist in the retina, lateral geniculate, and cortex – behave in an opponent fashion, being excited, for example, by green light in the surround of the neuron's receptive field and inhibited by red light in its center (see Zeki 1993; cf. also Hardin 1988).

Now, in our view, in order to correctly model the function of opponent neurons, one has to appeal to opponent process theory. That is, in the absence of the psychological theory, one cannot understand what opponent neurons do. Although the neurobiology of color opponency, therefore, might exhaust the mechanism of color opponency, it does not provide a complete theory. A complete theory of color opponency must appeal to the function of opponent neurons and the psychophysical framework of opponent process theory. A purely neurobiological theory of color opponency, therefore, does not exist even if opponent neurons are all there is to the mechanism of opponent color vision.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

Our claim about Kandel's theory of learning is analogous. Even if the synaptic mechanisms described in that theory are all there is to the mechanism of elementary learning, it does not follow that there is a purely neurobiological theory of elementary learning. Indeed, we claim that there is no such theory. In the sections that follow, we develop this argument by elaborating the idea that Kandel's theory is not purely neurobiological. We do this by considering a number of objections to the argument just described.

5.3.2. An objection concerning reduction. It might be objected that we have set an unreasonable standard for Kandel's theory, and, by implication, for any other putative neurobiological theory of a mental phenomenon. Let us suppose that Kandel's theory is genuinely explanatory of the neurophysiology of classical conditioning, and let us further suppose that though it appeals to psychological notions or theories, it can nonetheless explain conditioning in neural terms. Under these assumptions, shouldn't we say that the presence of psychology in the theory is harmless?

Consider a familiar analogy already mentioned. We know that temperature is mean kinetic energy, but we continue to refer to temperature nonetheless. The mere fact that we do so, however, does not mean that there is anything inadequate in the original identification. Similarly, one might argue that the presence of psychology does not affect the success of Kandel's theory at explaining conditioning in neurobiological terms. On this view, Kandel's theory is purely neurobiological in any sense that matters scientifically, and therefore it counts as support for the radical neuron doctrine.

5.3.3. Response: Reduction and implementation. We think this objection misconstrues the role of the psychology in Kandel's theory. In order to explain why, we distinguish two kinds of case in science in which an intuitively more basic theory is brought to bear on the phenomena explained by an intuitively less basic theory.

The first case is *reduction.* As we remarked above (sect. 4.2), reduction is the process whereby the concepts and laws of the more basic reducing theory are used to explain the phenomena described in the less basic reduced theory. Because the reduced theory can be derived from the reducing theory, the latter is conceptually independent of the former. The reduction of temperature to mean kinetic energy is such a case. Because the phenomena explained by the concept of temperature and its associated laws (such as they are) can be derived from the concept of mean kinetic energy and *its* associated laws, the kinetic theory can explain at least as much as the temperature theory can. In principle, therefore, the concept of temperature and its laws can be ignored without any loss of explanatory power.[41]

The second case is *implementation,* whereby a more basic theory provides the mechanistic details of the system that instantiates the functions posited by the less basic theory. The best-known illustration of implementation is Marr's (1982) conception of the role of neurobiology in the theory of vision. (For a more general discussion, see Cummins 1983.) According to Marr, the neurobiology of vision describes how a particular psychological process, which Marr referred to as an algorithm, occurs in a particular neural system. In turn, the algorithm is a particular instantiation of a more abstract computational process. In cases of implementation, the conceptual work is done by the computational theory, and not by the algorithm or the implementation. However important the neurobiological story is, it remains conceptually parasitic on the higher-level theory. If the computational level of the theory were eliminated, the mechanistic details would no longer make sense. As Marr (1982) famously remarked, it is impossible to explain how a wing works simply by describing its feathers.

Implementation can hold between neurobiology and psychology whether or not the psychological story is computational; other psychological accounts will do (Cummins 1983). The case of color opponency discussed earlier can be used to illustrate this point as well. We suggested above that the properties of color-opponent cells might constitute the mechanism that causes color vision to have an opponent character. One can therefore think of the physiology of color-opponent cells as implementing opponent process theory because, as we suggested, in the absence of that theory, one cannot explain color opponency or what these neurons do.

Nor is the relation of implementation restricted to theories of psychological function. Suppose, for example, that someone were to produce an explanation of gene replication in terms of particle physics. It is plausible that although such a theory would explain what is happening at the particle level during replication, the explanation of the basic phenomena of replication would continue to reside at the molecular level. In such a case, it would be appropriate to call the particle story an implementation of the molecular story rather than a reduction of it.[42]

Given this distinction, the question is whether Kandel's theory is an instance of reduction or implementation. We think there is good reason to suppose it is an instance of the latter. Hawkins and Kandel themselves say "Our goal is thus to suggest how cognitive psychology may begin to converge with neurobiology to yield a new perspective in the study of learning" (1984, p. 376).

Gluck and Thompson are more explicit: "Hawkins and Kandel (1984) have taken a formidable step in attempting to bridge the gap between algorithmic level models of classical conditioning and implementation-level models of the underlying neurophysiology" (1987, p. 189). These remarks express the view that the neurobiological theory of learning fills out the conceptual structure devised by psychology. It does not provide an independent explanation of the phenomena. Kandel's theory could not, therefore, stand on its own without appealing to psychological theory of some kind. It is conceptually parasitic on those theories and must be classed as implementation.

5.3.4. Rejoinder: Interpretation aside, doesn't Kandel's theory *in fact* provide a conceptually independent description? In response to our claims above, one might argue as follows.[43] Suppose one gave a neuroscientist a description of activity-dependent presynaptic facilitation in the classical conditioning paradigm. Perhaps the story might go like this: stimulus 1 produces a moderate response in the sensory neuron; it releases a small quantity of neurotransmitter, which fails to elicit the gill-withdrawal reflex. With repeated near-simultaneous activation of a connected pathway by stimulus 2, the response of the sensory neuron is enhanced, a greater quantity of neurotransmitter is eventually released, and its capacity to fire the motor neuron is facilitated; and so on. Wouldn't it be fair to say that our neu-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

roscientist understands classical conditioning as well as any psychologist *without* the need for psychological concepts? If so, then it looks as if the concept of activity-dependent presynaptic facilitation is not parasitic on the psychology of conditioning, and our claim that Kandel's theory is implementation rather than reduction is mistaken.

5.3.5. Response: The complexity of conditioning. The objection is tempting, on our view, only if one underestimates the psychological complexity even of the elementary forms of learning dealt with by the Kandel model. This becomes clear if we briefly examine the psychology of classical conditioning.

A pervasive view of classical conditioning, even in textbook presentations, is that it is a process whereby contiguity of the US and the CS transfers the response from the former to the latter. However, as Rescorla (1988) argues, although this view is adequate to the conception that Pavlov himself had – a conception that emerged from the reflex tradition – it is entirely inadequate to account for the data that have been accumulated during the last 20 years of research on conditioning.[44]

Consider, first, some relevant data. Contiguity is essential to the traditional theory and is supposed to be sufficient to elicit conditioning. In an early study, however, Rescorla (1968) compared two related learning situations. In the first situation, a rat is exposed to the CS, a tone that is presented randomly during a 2-minute interval. During the same interval, the US, an electric shock, is also randomly applied (Fig. 2A). In the second situation, the shock occurs only when the tone is presented (Fig. 2B). In the latter case, but not the former, an association between the tone and the shock is learned. Notice, however, that the CS and the US occur within the same overall period of tone and shock exposure, thus satisfying the requirement of contiguity. Nevertheless, only one pattern of CS and US pairing produces conditioning.

Of course, conditioning occurs in the latter case because the tone provides *information* about the occurrence of the shock. In order to explain this effect, therefore, one needs to appeal to some notion of information that is richer than the notion of "low-level mechanical process in which the control over a response is passed from one stimulus to another" (Rescorla 1988, p. 152).

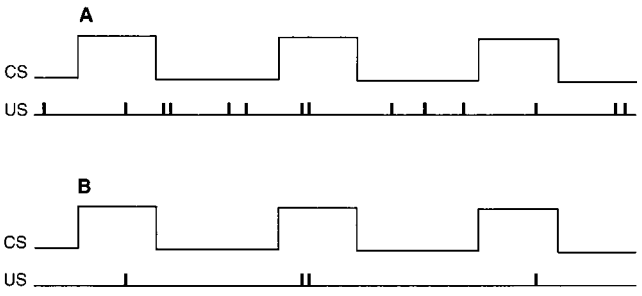


Figure 2. A classical conditioning paradigm demonstrating the insufficiency of contiguity for conditioning. In the first case (**A**), the conditioned stimulus (CS) is always paired with the unconditioned stimulus (US), but the US also occurs at other times during the training interval. In the second case (**B**), the US occurs only when the CS occurs. Although the requirement of contiguity is satisfied in both cases, conditioning is achieved only in the latter case. (Adapted from Rescorla 1988.)

Here is a second case (see Rescorla 1980). In a variation of classical conditioning called *autoshaping*, a bird learns to associate a red square (the CS) and food (the US), and will eventually come to peck at the square as if it were the food itself. The bird will also learn to peck at stimuli that are associated with the original CS – a form of learning known as *second-order conditioning*. Now consider two different conditions. In the first, an achromatic outline of a square is associated with the red square; in the second, an achromatic outline of a triangle is associated with the red square. In the first case but not the second, the second-order stimulus relates to the first-order stimulus as part to whole, and in the first case, conditioning occurs much more readily than in the second. A mental representation of a *relation* governing the stimuli thus has a differential effect on the course of learning (Fig. 3). Once again, the traditional conception of contiguity will not explain the effect.

In general, recent research has shown that classical conditioning is more complex in many respects than Pavlov and others believed (see Rescorla 1988). For example: (1) the context of learning is relevant; (2) the animal learns a variety of associations among many stimuli simultaneously in the learning situation; (3) associations exhibit a hierarchical organization; and (4) the response to the CS depends not only on the response to the US but on the properties of the CS itself: for instance, a tone signaling shock will cause a rat to freeze, but a prod signaling shock will cause the rat to try to cover the prod up. Classical conditioning is thus more than the transfer of response from US to CS. It requires positing, as Rescorla puts it, "the learning of relations among events that are complexly represented, a learning that can be exhibited in various ways" (1988, p. 158). Thus,

> the simple pairing of two events cannot be taken as fundamental to the description of Pavlovian conditioning. Instead, [these experiments] encourage the prevalent modern view that con-

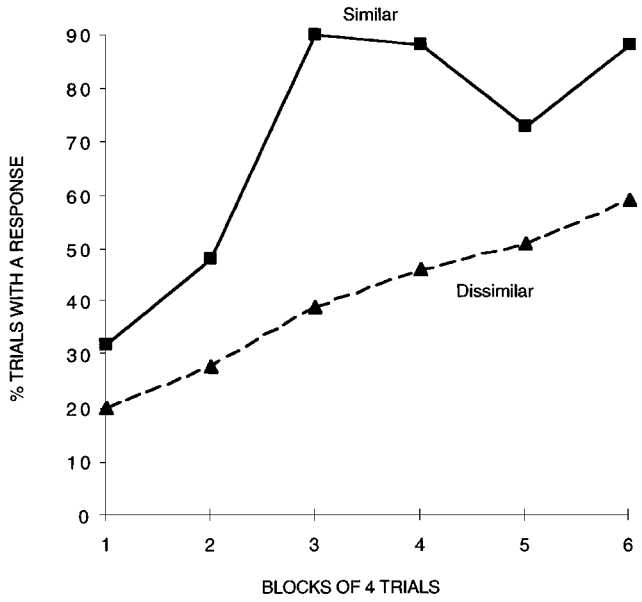


Figure 3. Learning an association between a second-order stimulus that bears a part-whole relation to a first-order stimulus. In the *similar* case, the second-order stimulus bears the part-whole relation to the first-order stimulus, and in the *dissimilar* case it does not. (Adapted from Rescorla 1988.)

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

> ditioning involves the learning of relations among events. It provides the animal with a much richer representation of the environment than a reflex tradition would ever have suggested. Of course, one cannot leave the analysis at this level; rather, one needs to provide theories of how these relations are coded by the organism. Such theories are now available, several of which are stated in sufficient quantitative detail to be taken seriously as useful accounts. . . . These theories emphasize the importance of a discrepancy between the actual state of the world and the organism's representation of that state. They see learning as a process by which the two are brought into line. In effect, they offer a sophisticated reformulation of the notion of contiguity. A useful short-hand is that organisms adjust their Pavlovian associations only when they are "surprised." (1988, p. 153)

This richer conception of classical conditioning presents a dilemma for the view that Kandel's model is a reduction of conditioning and the other forms of elementary learning. The first horn of the dilemma is this. The basic concept available to the Kandel model is the concept of synaptic plasticity – in particular, the process of activity-dependent presynaptic facilitation – but it is hard to see how this concept could capture the conceptual complexity either of the notion of information about relations or of surprise, among others. This is *not* to say that synaptic changes are not the mechanism of information or surprise; they may well be. Even if they are, however, the *concept* of synaptic change cannot capture the *concept* of information or surprise. Claiming that Kandel's model reduces classical conditioning in *Aplysia,* therefore, would simply ignore some of the concepts necessary to explicate the contemporary conception of conditioning. If Kandel's model *were* a reduction of conditioning, it would be empirically inadequate. Further, if one were to claim that conditioning in *Aplysia* in particular only requires the reflex conception and not the richer one, this would answer our objection at the cost of rendering Kandel's theory an explanation isolated from most of the rest of learning theory and irrelevant to learning in mammals, including humans.

The second horn of the dilemma is this. In response to the objection just made, one could argue that Kandel's model is neutral with respect to the reflex and the modern conceptions of classical conditioning, and indeed to any other conception. Kandel's theory could be interpreted as an instance of *whichever* notion turns out to be correct. On this line of argument, however, Kandel's model cannot be a reduction of the theory of conditioning because it is neutral with respect to incompatible theoretical notions. Because it does not tell for or against a particular theoretical conception, it cannot reduce any particular one. To take an analogy: a putative reduction of optics that did not decide for or against the wave conception of light or the particle conception (or a mixed conception) could not successfully reduce optics.[45] Similarly, no putative reduction of classical conditioning that speaks neither for nor against some essential theoretical characterization of conditioning can be successful.

We conclude that the claim that Kandel's model is a reduction of classical conditioning, rather than an implementation of it, cannot be sustained because it is not possible to understand the full complexity of conditioning without recourse to some psychological theory or other. If Kandel's model is accurate, it can only be a representation of how the relevant psychological notions are instantiated in the neural machinery.

We have only considered classical conditioning here, but we suspect that similar remarks could be made about other forms of elementary learning (see, e.g., Wagner & Pfautz 1978), and if such remarks could not be made, this in itself would be problematic for Kandel because it would reintroduce theoretical variation (say, between conditioning and sensitization) where Kandel has argued for theoretical unity. What is true of conditioning had better be true of the other forms of elementary learning, or one of the significant virtues of the Kandel model is lost.

5.3.6. Summary. How does this discussion affect the question of the neuron doctrine? The radical neuron doctrine says that a successful theory of classical conditioning, among other phenomena, will be neurobiological. However if our argument is sound, then even if we consider Kandel's theory of elementary learning to be a correct neurobiological description of the phenomena, it does not follow that a successful theory of conditioning will be solely neurobiological, any more than it follows from the implementation of color vision by color-opponent cells that a successful theory of color vision will be neurobiological, or from the imaginary particle physics description of replication that a successful theory of replication will be a part of physics. What determines the form of a successful theory is where the best explanation is to be found, and in the present case we take the best explanation to reside with psychology and not neurobiology because we take the latter to be parasitic on the former. We think that it is currently an open question what form a successful theory of learning will take and that the answer to this question will depend in part on whether neurobiology can produce conceptually independent accounts that are superior to those already to be had in psychology. Because Kandel's theory is not purely neurobiological, therefore, appealing to it in the context of the argument from exemplars does not support the radical neuron doctrine.

5.4. Is Kandel's model a special case?

Our interest in Kandel's theory of learning in *Aplysia* has been driven by the idea, suggested in the passage we quoted from Hubel, that exemplars of neuroscientific theorizing might be taken as an inductive basis from which to infer the radical neuron doctrine. We have argued that although Kandel's account of learning is an important part of neuroscientific theory, it does not support an inference to the radical doctrine because it involves psychological notions.

One might respond to our argument with the suggestion that Kandel's theory is in some sense a special case and that there are other cases of successful neuroscientific theory that might do the job better. There are a variety of senses in which one could take the phrase "special case." We consider three and try to show that Kandel's theory is not a special case in any sense that matters to the argument from exemplars.

5.4.1. The problem of a single case. One might argue that, given its structure, the instance of the argument from exemplars supporting the radical neuron doctrine is not refuted by the failure of a single case. The argument requires that there be *some* exemplars that support the radical neuron doctrine, and the failure of this particular case does not show that there are no others that would do the job. Indeed,

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

one might argue, for the same reason, that the argument would not strictly be refuted by the failure of a number of cases.

These claims are quite true, but they are irrelevant to our critique because they apply to *all* arguments of the present form and not to ours in particular. We cannot prove that there are, or will be, no cases that support the argument from exemplars. A reasonable agnosticism about the future of science explicitly prohibits us from trying. However, the burden is on the supporters of the argument to offer a better candidate than Kandel's theory. In the absence of such a candidate, one ought to accept that the argument fails to support the radical neuron doctrine even though it has not been strictly refuted.

5.4.2. The problem of an inadequate case. A second way in which Kandel's theory might be claimed to be a special case is if it were to be shown that the theory were not the best neuroscience could offer. If there were other neuroscientific theories of mental phenomena that answered the requirements of the argument from exemplars, our choice of Kandel as a paradigm case would be unfair to the defender of the radical neuron doctrine.

To this we can only appeal to the consensus in the neuroscientific community that we mentioned at the outset of our discussion. In our opinion, Kandel's theory is widely recognized as the sort of model to which neuroscience aspires and, to that extent, is not special in the sense of being a poor representation of neuroscientific theory.[46] We know of no other theory that is as successful as Kandel's, but we are, of course, prepared to be convinced otherwise.

5.4.3. The problem of a case with unique features. There is one important sense in which it might be argued that Kandel's theory is a special case. If one could show that the theory had unique features that supported our critique above – that is, features that supported our claim that psychological theory is necessary to interpret it – then one could argue that our view of Kandel does not defeat the argument from exemplars. That is to say, if it could be shown that future neuroscientific theories will not rely on psychological notions as Kandel's theory does, then the argument from exemplars could be made successful by appealing to the conjunction of the scientific success of Kandel's theory *and* this ancillary claim of the independence of future theories from psychology.

We have suggested that mental phenomena must currently be addressed by means of broadly psychological theories and that neurobiology is not now in a position to offer competing theories. In responding to this objection, it suffices for us to note that Kandel's theory deals with extremely simple forms of behavior such as classical conditioning. Even here, a highly sophisticated, mathematically rigorous psychological research program of more than a century has not exhausted or resolved all of the theoretical questions. With more complicated phenomena, it seems very likely that more, as well as more elaborate, psychology will be necessary. This is only what one would expect; it is practically impossible to tie a neurobiological theory to a psychological phenomenon without a detailed fractionation of that phenomenon in psychological terms. It would be absurd, for example, to ask for a neurobiological account of learning without providing at least a taxonomy of kinds of learning, and, more importantly, a detailed psychological story about how some type of learning is supposed to occur. For without such a detailed story, how would the neurobiologist recognize which facts potentially explain the phenomena? As the phenomena to be explained get more complicated, things can only get worse from the point of view of the radical neuron doctrine: more psychology, not less. One hopes and expects that neurobiology will modify psychological theories, but there is absolutely no evidence that it has any genuine alternative to offer. Neurobiology may one day invent concepts that can compete with psychology, but that day has not yet come. Hubel puts it this way: "A revolution of Copernican proportions has not yet occurred in neurobiology, and will perhaps not occur, at least in a single stroke. If there is one, it may be gradual, taking place over many decades. When it is over, we will know whether the brain is capable of understanding the brain" (1974, p. 259).

Finally, although we have suggested that one cannot interpret Kandel on learning as providing support for the radical neuron doctrine, we have not meant to be criticizing Kandel in the least. For one thing, as we have noted, Kandel's theory contributes to our conceptual understanding of learning. In arguing, for example, that the mechanism of conditioning develops out of the mechanism of sensitization, Kandel is making a substantive claim about learning that, if true, may change psychological taxonomy. It is this sort of contribution that represents what neuroscientists and philosophers expect neuroscience to offer in the future: a contribution to the way we think about the basic phenomena of the mind. However, this may fall well short of a wholesale replacement of our concepts.

Moreover, although we do not believe that if one were to remove all of the psychology from Kandel's theory of learning there would be an adequate theory left standing, we do not regard it as a criticism, but rather a virtue, of the theory that it draws on explicitly psychological notions. Psychological theory is currently necessary both in the discovery of the biological facts and in their interpretation, and this is particularly true in the domains where psychological theory is detailed and successful. In the absence of a revolutionary new neurobiology, it seems to us that the integrative approach of cognitive neuroscience is the smart bet.

6. Conclusion

We have argued that a very common view in neuroscience and philosophy is subject to two interpretations, one trivial and one radical, and that initially persuasive lines of argument in favor of the radical view fail. What conclusions should one draw from our discussion?

The most important conclusion concerns the proponents of the neuron doctrine. In response to the ambiguity in the doctrine that we have pointed out, proponents of the doctrine are faced with a dilemma. On the one hand, they might respond that what they intended to defend was only the trivial doctrine. Although this is a perfectly reasonable response to our argument, it means, as we noted in section 2, that proponents of the doctrine are in the position of defending a scientific triviality. On the other hand, proponents of the doctrine might bite the bullet and adopt the radical version of the view. The trouble with this position, however, is that it is not clear that there are any good arguments for it. Proponents of the radical neuron doctrine are therefore in the uncomfortable position of holding a scientific view

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

for which no obvious scientific justification is available. Of course, one might imagine a nonscientific argument for the radical doctrine, but in our view that would be perverse and anyway has not been given. As we remarked in section 2, therefore, it is incumbent on defenders of the neuron doctrine to explain why they are defending a view that either has no defense or that needs none.

In addition to this general moral, there are two more specific morals that are suggested by our article, one for neuroscience and one for philosophy. We close with a brief discussion of them.

6.1. Neuroscience: The invisible hand in the neuron doctrine

The first moral concerns the practice of neuroscience itself. The annual American neuroscience meeting hosts approximately 25,000 scientists, and the amount of work presented there is commensurate with those numbers. However, despite the enormous data we are accumulating about the brain, broad theoretical suggestions are few and far between. There may be various reasons for this conservatism in neuroscience, and we are not in a position to make confident judgments about this, but we have one speculation to offer. Neuroscientists may be reluctant to theorize about the mind because they have an unreasonably strict view about the sort of theory that is "properly neuroscientific." A neuroscientist with a relatively stringent understanding of the neuron doctrine might hold a similarly stringent view about the conceptual resources available to theory, shying away from explicitly incorporating psychological theories, or theory fragments, into what is supposed to be a neuroscientific account.

If our argument is sound, however, then this self-imposed restriction is a mistake – an effect of the invisible hand of confusion about the neuron doctrine. For this reason, we are encouraged by the beachhead that cognitive neuroscience has begun to establish in contemporary neuroscience.

6.2. Philosophy: A plea for more philosophy of neuroscience

The second moral concerns the development of a philosophy of neuroscience. Among the various sciences that study the mind and brain, neuroscience is perhaps the most widely held to have philosophical import. There is frequent reference to neuroscience in the philosophy of mind literature, and neuroscientists themselves often suggest that their discipline has relevance to the problems discussed by philosophers. As Zeki (1993) says, "ultimately the problems that cortical neurobiologists will be concerned with are the very ones that have preoccupied the philosophers throughout the ages." So it is a curious fact that there is relatively little philosophical discussion of the basic concepts and theses of neuroscience. Although physics has led to a philosophy of physics, and biology to a philosophy of biology, there is no philosophy of neuroscience, and this is a lacuna in the philosophy of science.

By "philosophy of neuroscience" we mean something distinct from what has come to be called *neurophilosophy* (P. S. Churchland 1986). Neurophilosophy, as it is commonly understood, is an application of the results of neuroscience to problems in the philosophy of mind. In contrast, we take philosophy of neuroscience – by analogy with philosophy of physics and philosophy of biology – to be concerned primarily with the presuppositions and philosophical problems of neuroscience itself. Philosophy of neuroscience and neurophilosophy may interact, but they are distinct enterprises.

Part of our aim in this article has been to make a contribution to the development of the philosophy of neuroscience. Whatever the value of the present effort, we are convinced that a close investigation of the foundations and concepts of neuroscience will play a part in the progress of neuroscience and in the development of the science of the mind which the millennial future will see realized.

ACKNOWLEDGMENTS

Versions of this article were read at the Australasian Association of Philosophy Conference, 1996; in the Philosophy Program at the Australian National University; in the Cognitive Science Group at the Australian National University; in the PNP program at Washington University, St. Louis; and in the Department of Philosophy at the California Institute of Technology. We are indebted to members of the audiences on those occasions for very helpful discussions. We would also like to express our gratitude to Bill Bechtel, David Braddon-Mitchell, Noam Chomsky, Hugh Clapin, Michael Cook, Marilyn Friedman, Steve Gardiner, Jay Garfield, Brian Garrett, David Hilbert, Frank Jackson, Christian Perring, Walter Sinnott-Armstrong, Michael Smith, Natalie Stoljar, and Louis Taillefer for their comments on earlier versions of this paper. We are especially grateful to five *Behavioral and Brain Sciences* referees – Max Coltheart, Gilbert Harman, James Higginbotham, John Kihlstrom, and one anonymous reviewer – for their many constructive criticisms and suggestions that helped us to improve the article considerably.

NOTES

1. The phrase "neuron doctrine" originally referred to a view, formulated in 1891 in a famous review by Wilhelm Waldeyer, expressing the upshot of the seminal work of the anatomist Santiago Ramón y Cajal (although the doctrine was developed as a result of the work of many individuals, notably Camillo Golgi). This work revealed that the brain, like the rest of the body, is made up of cells. More specifically, the doctrine expressed the view that "the nerve cell is the anatomical, physiological, metabolic, and genetic unit of the nervous system" (Waldeyer quoted in Shepherd 1991, p. 4). The formulation of the neuron doctrine marked the resolution of a dispute between those who believed that neurons were bounded entities and those who believed, in accordance with the opposing *reticular* theory, that the fine branch-like structures seen in the brain are continuous with one another. The doctrine thus entailed that neurons must communicate with one another by contact. The neuron doctrine represents the fundamental tenet of modern neuroscience and is a claim about the brain and its function. In this article, however, we use the phrase "neuron doctrine" to refer to a view about the relation between the neural and the psychological or mental. We are not, however, doing the phrase as much violence as one might suppose. G. M. Shepherd (1991), one of the modern historians of the doctrine, writes in his introduction to the history of the subject: "Of broader interest is the potential significance of the neuron doctrine as one of the great ideas of modern thought. One thinks here for comparison of such great achievements of the human intellect as quantum theory and relativity in physics; the periodic table and the chemical bond in chemistry; the cell theory, evolution, and the gene in biology. Notably missing from this register is a theory for explaining how the brain makes these accomplishments and all other human activity possible. The pioneers of the neuron doctrine believed that they were laying the foundation upon which such a theory had to be built. Descartes had set the philosophical agenda for the mind – body problem some 300 years previously, but these scientists were

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

the first to come face to face with the cells and their connections where that problem will likely have its resolution" (pp. 9–10).

2. There could also be probabilistic versions of the doctrine according to which it is likely or possible or probable that a successful theory of the mind will be solely neurobiological. For clarity and simplicity, we will for the most part ignore these versions here, and concentrate on the simple formulation of the doctrine.

3. In section 3 we consider whether this line of argument is successful.

4. The thesis of materialism (or physicalism) is often stated in a manner that is more sophisticated than the formulation we use in the text; for example, it is often stated as a *supervenience* thesis. However, the simpler statement will do for our purposes. It is also worth noticing that although materialism is usually taken for granted, and although we ourselves will assume here that it is true, there are certainly people who deny it. For a very good recent discussion of materialism (and for an argument that it is false), see Chalmers (1996).

5. For philosophy of language aficionados: the qualifier "practically" is required because from "$A = B$" it does not follow that "the science of A = the science of B"; that is, "the science of" is an intensional functor.

6. For a discussion of the impact of the neuron doctrine on domains further afield, see Churchland (1995). Here we concentrate only on the scientific consequences of the doctrine.

7. For evidence of the esteem in which the Churchlands are held, see McCauley (1996).

8. As one of our *Behavioral and Brain Sciences* referees put it, the Churchlands "only say boldly what a lot of other neuroscientists say *sotto voce*."

9. In a discussion of the dimensions along which facial recognition occurs in the brain, for example, Churchland confesses that "it's not known exactly what those dimensions are, nor even that they are identical in all of us" (1995, p. 19). A similar point is made by Fodor (1998, p. 84).

10. The reason for the qualification concerns the enormous difficulty involved in explaining the conscious aspects of perception: "In 1972 , I suggested that 'active high-level neurons directly and simply cause the elements of our perception,' and I still think this simple idea has some merit, even though I now believe that interactions with other individuals and society have to be taken into account when considering the conscious aspects of perception" (Barlow 1995, p. 428). For further discussion of this issue, see Barlow 1987.

11. She speaks, for example, of the "functionalist research ideology" promoted by Jerry Fodor and others (see P. S. Churchland 1986, Ch. 9 passim).

12. It is important to note that Hubel's view is not exhausted by the passage above, as is evident if one looks at the passages we discuss below.

13. One might distinguish the notion of being a second-grade science in the fullness of time and being a second-grade science at the moment. It is possible that science must go through a non-neuroscientific stage of explanation in order to arrive at a neuroscientific explanation, thus making linguistics, for example, first-grade science for the moment. Our concern is with scientific explanation in the fullness of time, and we ignore the above distinction for the sake of simplicity. See also the discussion of reductionism in practice versus reductionism in principle in note 30.

14. It is precisely this picture that one finds in descriptions of classical psychology such as Cummins (1983). That the trivial doctrine does nothing to alter this picture is one reason why it is trivial.

15. It is worth emphasis that we are not claiming that the trivial neuron doctrine is trivial in the logical sense, nor do we mean that those who deny the trivial version are making an obvious mistake, or a mistake of logic. Rather, in using the word "trivial" we intend to emphasize that the trivial neuron doctrine (a) is a doctrine that has, or ought to have, extremely wide support in the scientific community; and (b) does not have the consequence for the psychological sciences that the radical version of the doctrine does.

16. It is important to notice, however, that the trivial neuron doctrine is compatible with the view that the successful theory of the mind will be a solely neurobiological theory – the claim we identify below as the radical neuron doctrine. The trivial neuron doctrine is compatible with the radical neuron doctrine because a theory of the mind expressed solely in terms of neurobiological concepts is one version of a scientific theory of the brain. In addition, the trivial neuron doctrine is consistent with a theory of the mind expressed solely in terms of the concepts of psychology alone, and of course the trivial doctrine is consistent with any theory of the mind that is a combination of psychology and biology, or indeed physics, chemistry, and the other sciences, should they turn out to be relevant to understanding the brain. The trivial neuron doctrine is thus an extremely weak view; it expresses little more than a commitment to an explanation of the mind by science, presumably the sciences currently involved in its investigation.

17. One might object here that the radical neuron doctrine that we are considering is extremely strong, and that one can imagine versions of the doctrine that would be weaker but would nevertheless not simply be the trivial doctrine. For example, as we have defined it, the radical doctrine has the resources only of neuroanatomy, neurochemistry, and neurophysiology. However, could one not add *other* branches of neural science (or of science more generally), and, in consequence, weaken the neuron doctrine without it collapsing into the trivial doctrine? Although this is certainly a possibility, it is irrelevant to the main point we want to make. As we argue later, the explanations offered by neurobiologists of even elementary psychological processes are in fact not purely neurobiological and must draw on explicitly psychological theory. If this is right, however, the only way currently available to enrich biological neuroscience so as to explain the mind is to enrich it with psychology. However, this is in effect to adopt the trivial form of the neuron doctrine. Of course, this is not to claim that there is absolutely no possibility of neurobiology explaining the mind; it is only to insist that, as things currently stand, there is no evidence of that coming about.

18. We understand from Christof Koch (personal communication), however, that Crick is in fact a defender of the trivial neuron doctrine.

19. To be fair, we should point out that immediately after this passage, Churchland makes the following parenthetical remark: "The same goes, of course, for the revisions and reconstructions in neuroscience" (p. 374). Nevertheless, it is difficult to view this remark as detracting from the radicalness of the passage quoted in the text. The last sentence of the passage in the text is conditional, so the interpretative question is whether the parenthetical remark is intended to qualify the antecedent or the consequent of this conditional (or both). If it is intended to qualify the antecedent – namely, the claim that "psychology is willing to test and revise its theory and hypotheses when they conflict with confirmed neurofunctional and neurostructural hypotheses, and as long as the revisions are made with a view to achieving concord with a lower-level theory" – Churchland appears to be saying that so long as both psychology and neuroscience are willing to revise and test their hypotheses in the light of the each others' results, then psychological capacities and processes will finally be explained in neuroscientific terms. In other words, she is saying that an inevitable result of the "co-evolutionary strategy" is the radical neuron doctrine. However, in light of the evident logical gap between the co-evolutionary strategy and the radical neuron doctrine, it seems reasonable to view Churchland here as failing to make this crucial distinction between the two versions of the doctrine. On the other hand, if the remark is intended to qualify the consequent of the conditional – viz., the claim that "the capacities and processes described by psychological theory will finally find explanations in terms of neuroscientific theory" – then Churchland is saying that the capacities and processes described by neuroscientific theory will finally find explanations in terms of psychological the-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

ory. However, to interpret her this way is, first, to interpret her as saying something that is quite antithetical to the surrounding text, and second, does nothing to support the radical neuron doctrine.

20. It is perhaps worth mentioning here that there are other trends in the Churchlands' work that seem to represent some confusion about the neuron doctrine. One of these is what they see as a "monolithic" approach to the different levels of explanation found, they say, in Marr: "Marr's three-level division treats computation monolithically, as a single kind of level of analysis. Implementation and task-description are likewise each considered a single level of analysis" (Churchland & Sejnowski 1992, p. 19). Such a monolithic picture is objectionable because it rules out the idea that there are many levels of explanation for psychological and neural phenomena. The rejection of that monolithic picture seems to us quite right, but it is clear that rejecting the picture does not provide support for the radical neuron doctrine. If there are *many* different levels of explanation, then as long as the psychological sciences concern themselves with at least some of these, we have reason to believe only the trivial doctrine. The second trend in the Churchlands' work that represents an unclarity about the neuron doctrine is their commitment to connectionism (see, e.g., Churchland 1995). Connectionism is often thought to provide a way of theorizing about the mind that appeals only to neuron-like entities. As such, it might be tempting to suppose that connectionism might support the radical neuron doctrine. However, any such line of argument seems to rest on a confusion about what connectionism is. As we understand it, connectionism can be understood either as a very general implementation structure for psychological processes or as a proposal about the form of psychological theory itself. Whether connectionist theories could be empirically adequate on either interpretation is obviously a controversial matter, but the important point for us is that on neither interpretation does connectionism support the radical neuron doctrine because on either interpretation, connectionist approaches depend on psychology. For important discussions of connectionism, see Fodor and Pylyshyn (1988) and Pinker and Prince (1988).

21. There are a number of reasons for supposing that an argument of this style is in operation in the Churchlands. One is "the articulation of biological organization" passage that we have already quoted and commented on. Another is given by passages such as the following from Churchland and Sejnowski (1992): "The venerable old paradigm depicted humans as blessedly perched on the apex of the Great Chain of Being, lucky to have been created in the image of God, and fitted out with a nonphysical immortal soul housing a freely exercisable will, a consciousness that experienced feelings and sensations, and a rational faculty that mercifully could transcend the merely mundane, for example by proving mathematical theorems. The old paradigm was frankly supernaturalistic. It exhibited both species chauvinism in quite spectacular degree and a profoundly nonempirical acceptance of nonphysical forces, stuffs, and mechanisms. . . . The new paradigm is naturalistic, and it is shaped by the scientific image. By pulling out the linchpin assumption that humans are set apart from the natural order, it changed everything. The naturalistic approach to the mind-brain, foreshadowed by Hobbes and de La Mettrie in the seventeenth century, became a live possibility in the nineteenth, largely by dint of advances of microscope and staining technology, a nonoccult understanding of electricity, and the commanding scientific leadership exemplified in the breadth and depth of success of physics and chemistry. The pioneers were mind/brain scientists, especially du Bois-Reymond, Helmholtz, Cajal, Golgi, Jackson, and Wertheimer and the massive backdrop against which . . . the naturalistic vision made sense was Darwin's perspective on the origins of biological complexity. Although essentially constant in its ultimate goal, naturalism has been revived by recent discoveries in neuroscience and by a growing confluence with the computational and behavioral sciences" (1992, p. 142). In this passage, the references to Hobbes and de La Mettrie and to naturalism obviously suggest something like the argument from naturalism and materialism. Moreover, the emphasis on microscopes and staining technology strongly suggests that the argument is an argument for the radical neuron doctrine. Nevertheless, the final refrain in the passage leaves open the possibility that the doctrine under discussion is only the trivial doctrine.

22. We might also mention a third kind of objection here, deriving from *externalism* in the philosophy of mind. The lesson of externalism is often taken to be that the individuation of psychological phenomena depends crucially on the social, physical, or historical environment of the organism. If this is so, however, then one might think that the materialism expressed in premise 2 is too simple: psychological phenomena are not identical solely to neural phenomena. We discuss the pros and cons of this suggestion briefly in Stoljar and Gold (1998).

23. The phrase is Ned Block's.

24. Patricia Churchland (1986) puts it this way: "The unity of science is advocated as a working hypothesis not for the sake of puritanical neatness or ideological hegemony or old positivistic tub-thumping, but because theoretical coherence is the 'principal criterion of belief-worthiness for epistemic units of all sizes from sentences on up.' Once a theory is exempt from having to cohere with the rest of science, its confirmation ledger is suspect and its credibility plummets. To excuse a theory as *hors de combat* is to do it no favors" (p. 376; the quotation in the passage is from P. M. Churchland 1980). That both Churchlands place such a premium on the unity of science suggests that they might be sympathetic to the argument from unification. Note, however, that our argument is not meant to be an exegesis of their views on unification.

25. Modern biology is founded in part on the fundamental idea, known as the *cell theory,* that all living things are made up of cells. What makes modern neurobiology a branch of modern biology, therefore, is Cajal's neuron doctrine because it is that doctrine that established that the brain is also made up of cells (see Stoljar & Gold 1998).

26. We have taken the liberty of altering Maudlin's phraseology somewhat.

27. Maudlin cites an apposite passage from Gleick's (1992) biography of Richard Feynman to illustrate his point: "When a historian of particle physics pressed him [Feynman] on the question of unification in his Caltech office, he resisted. 'Your career spans the period of the construction of the standard model,' the interviewer said. . . .

'The standard model, standard model,' Feynman said. 'The standard model – is that the one that says that we have electrodynamics, we have trivial interaction, and we have strong interaction? Okay. Yes.'

The interviewer said, 'That was quite an achievement, putting them together.'

'They're not put together.'

'Linked together in a single theoretical package?'

'No.'

. . . 'What do you call $SU(3) \times SU(2) \times U(1)$?'

'Three theories,' Feynman said." (Gleick 1992, p. 433).

28. The Churchlands (1994) provide a clear description of what reduction involves: "[G]enuine reduction, when you can get it, is clearly a good thing. It is a good thing for many reasons, reasons made more powerful by their conjunction. First, by being displayed as a special case of the (presumably true) new theory, the old theory is thereby *vindicated,* at least in its general outlines, or at least in some suitably restricted domain. Second, the old theory is typically *corrected* in some of its important details, since the reconstructed image is seldom a perfect mirror image of the old theory, and the differences reflect improvements in our knowledge. Third, the reduction provides us with *deeper insight* into, and thus a *more effective control* over, the phenomena within the old theory's domain. Fourth, the reduction provides us with a *simpler* overall account of nature, since apparently diverse phenomena are brought under a single explanatory umbrella. And fifth, the new

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

and more general theory immediately *inherits all the evidence* that had accumulated in favor of the older theory it reduces, because it explains all the same data" (Churchland & Churchland 1994, p. 48).

29. The reduction-elimination continuum is also a continuum of radicalness; the more psychology is eliminated rather than reduced, the greater is the revolution brought about by neurobiology. Even in the case of complete elimination, however, the successful theory of the mind is still recognizable as neurobiology rather than as some entirely new science.

30. To avoid confusion, however, it is perhaps worth mentioning and setting aside two complications that might be thought to have a bearing on our argument. The first complication concerns the difference between *metaphysical* and *explanatory* interpretations of reductionism. On the metaphysical interpretation, the reductionist is understood as saying that psychological phenomena are nothing more than neural phenomena. On the explanatory interpretation, the reductionist is understood as saying that *the theory of* psychological phenomena is reduced to *the theory of* neural phenomena – that is, that psychology is reduced to neurobiology. It should be clear that the thesis under discussion in the text is reductionism in its explanatory guise. As we have already seen, the fact that *A*s are made up of *B*s does not entail that *A*s are to be explained in terms of *B*s. Similarly, the fact that psychological phenomena are reduced in the metaphysical sense to neural phenomena does not mean that the former are to be explained in terms of the latter. The second complication concerns the distinction between reduction *in principle* and reduction *in practice.* These are clearly different doctrines: to deny that one theory is reducible in practice to another theory is not to deny a reductionism in principle. We are here discussing reduction in principle. If premise 2 of the argument from unification were interpreted simply as urging a reduction in *practice* of psychology to neurobiology, then it would seem highly implausible, given the impoverished state of current neurobiology.

31. Ned Block has a appropriate name for the form of argument employed here. He calls it "the reductionist cruncher." See Block (1990), pp. 279–80.

32. Some of the other theories we think might be worth investigating as putative neurobiological exemplars include: auditory perception in the barn owl (Konishi 1995); perception in the electric fish (Heiligenberg 1991); receptive field explanations of visual behavior (Hubel 1988); cone explanations of color-matching behavior (King-Smith 1991); magno- and parvocellular explanations of visual phenomenology (Livingstone & Hubel 1987); neural population models of motor planning (Georgopoulos et al. 1989); hippocampal function in learning (e.g., Buzsáki 1989); the 40-Hz oscillation hypothesis concerning perceptual binding (Gray et al. 1989); and the activation-synthesis theory of dreams (Hobson 1990). In Stoljar and Gold (1998), we consider the case of long-term potentiation (Bliss & Lømo 1973).

33. The most famous theoretical description of neural plasticity is that of Donald Hebb (1949), who proposed that functionally connected neurons that were active simultaneously would have their functional connection strengthened. This principle has come to be called *Hebb's Rule.* [See also Amit: "The Hebbian Paradigm Reintegrated – Local Reverberations" *BBS* 18(4) 1995; Pulvermüller: "Words in the Brain's Language" *BBS* 22(2) 1999.]

34. Kandel himself seems to think that the model will generalize to the class of learning phenomena called *implicit learning* (see Bailey & Kandel 1995).

35. Thus, for example, P. S. Churchland (1986): "At the cellular and molecular level Kandel and his colleagues . . . have discovered much concerning the neurobiological basis of habituation, sensitization and classical conditioning in the invertebrate *Aplysia californica.* . . . These discoveries are truly remarkable both because they represent a landmark in the attempt to understand the neurobiological basis of plasticity and because they show that memory and learning can, despite the skepticism, be addressed neurobiologically" (p. 369).

36. But see section 5.3.5 for a critique of this definition of classical conditioning. For a more detailed description of these and other forms of learning, see Manning and Dawkins (1992).

37. It is interesting to note that Kandel's theory thus supports Barlow's views mentioned above about the role of single neurons in neuroscientific explanation.

38. See DeZazzo and Tully (1995) for a related phenomenon in *Drosophila.*

39. In his classic book on the history of psychology, for example, Boring (1950) writes that Watson "adopted the conditioned reflex of Pavlov as the behaviorist's substitute for association" (p. 644).

40. It is worth noting that, even in physiology, the notion of reflex is highly theoretical. Charles Sherrington, one of the early developers of the concept of reflex, wrote: "A simple reflex is probably a purely abstract conception, because all parts of the nervous system are connected together and no part of it is probably ever capable of reaction without affecting and being affected by various other parts, and it is a system certainly never absolutely at rest. But the simple reflex is a convenient, if not probable, fiction" (quoted in Posner & Raichle 1994, p. 5). See also Clarke and Jacyna (1987): "The evolution of the concept of reflex activity, and its background of sensorimotor physiology . . . have received more attention from writers than any other topic in the history of the neurosciences" (p. 101).

41. See note 28 for the Churchlands' description of the virtues of reduction.

42. Because we are concerned with the general relation of implementation and not with computational theories alone, we do not consider what, if anything, plays the role of algorithm and computational theory in Kandel's model.

43. We owe this way of putting it to Max Coltheart.

44. Throughout this discussion, we rely on Rescorla (1988). See also the references therein.

45. It *could,* of course, *eliminate* the wave, particle, and mixed conceptions.

46. For example, a recent *New York Times* article (Hall 1998) says that "[n]o one has dominated this area of research more than Eric Kandel" (p. 30) and that "[m]any people share the view of Larry Squire, a leading neuroscientist at the University of California at San Diego, who calls Kandel's lifetime devotion to the study of memory a 'monumental' achievement" (p. 30). In a different popular article about Kandel's work (Touchette 1996), an MIT researcher is quoted as saying, "He drives the field by providing an intellectual structure that can be tested" (p. 35). Another researcher is quoted as saying, "Someone presents a model until there is enough evidence to either support or refute it. If you present your model first, you're king of the mountain until someone else comes along to knock you off"; the writer adds, "And for several decades, Kandel has been king" (p. 35).

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Biological neuroscience is only as radical as the evolution of mind

Terry Blumenthal and James Schirillo

Departments of Psychology and Neuroscience, Wake Forest University, Winston-Salem, NC 27109. **{blumen;shirija}@wfu.edu**

Abstract: A biological neuroscientific theory must acknowledge that the function of a neurological system is to produce behaviors that promote survival. Thus, unlike what Gold & Stoljar claim, function and behavior are the province of neurobiology and cannot be relegated to the field of psychological phenomena, which would then trivialize the radical doctrine if accepted. One possible advantage of adopting such a (correctly revised) radical doctrine is that it might ultimately produce a successful, evolutionarily based, theory of mind.

Gold & Stoljar (G&S) propose a radical neuron doctrine stipulating that neurophysiology, neuroanatomy, and neurochemistry – in essence, biological neuroscience – will by themselves eventually provide the conceptual resources to understand the mind. G&S present this doctrine as radical in that it operates outside the framework of function and behavior, which are relegated to the psychological sciences. However, their preconception of biological neuroscience is too limited. Neural systems have evolved to perform various functions, those functions that promote the biological survival of the species. Thus, to understand a neural system requires understanding its behavioral output. So, when G&S claim that "a successful theory of the mind will be a theory of the brain expressed in terms of the basic structural and functional properties of neurons, ensembles, or structures" (sect. 2.2, para. 3), they must also accept that this requires understanding the behaviors that result from such neural processes. These behaviors will either be continued or be eliminated in future generations owing to the pressures of natural selection, which ultimately dictate the types of neuronal systems under consideration. Consequently, the neural system of an individual organism is dependent on the prior behavior of the species, which was determined, in part, by the neuronal systems of individuals of that species. Thus, especially in biological neuroscience, it is imperative to tie organic processes to the behaviors of the organism involved. This is the same criterion that G&S incorrectly reserve for mental phenomena.

A solely biological neuroscientific theory requires accepting that the function of a neurological system is to produce behaviors that allow the species to reproduce, because this is why a specific neurological system exists. This notion deflates the proposed radical theory. In contrast, G&S claim that "neurobiology has no concepts that can be used to describe the behavior of an animal" (sect. 5.3.1, para. 2). For example, they claim that "a complete theory of color opponency must appeal to the function of opponent neurons and the psychophysical framework of opponent process theory. A purely neurobiological theory of color opponency, therefore, does not exist even if opponent neurons are all there is to the mechanism of opponent color vision" (sect. 5.3.1, para. 4). However, to understand such a neural organization requires an understanding of what "opponent neurons do." This *is* the province of neurobiology and cannot incorrectly be relegated to the field of psychological phenomena, which would then trivialize the radical doctrine if accepted.

The claim that "psychological theory is currently necessary both in the discovery of the biological facts and in their interpretation" (sect. 5.4.3, para. 5) is, therefore, unnecessary. Neurobiological science already requires those aspects of psychological theory that G&S claim weaken the doctrine, namely, that neurobiology is a functional process and that understanding the behavior that results from such systems is necessary to any study of neurobiology. However, there is one possible advantage to adopting such a (correctly revised) radical doctrine. We agree with the G&S statement that "the claim that the theory of mind will be expressed in cognitive neuroscientific terms expresses nothing more, therefore, than an ecumenism in the development of the theory and an agnosticism about its content" (sect. 2.2.1, para. 2). Insofar as the radical doctrine is predicated upon evolution and natural selection, however, it is hardly agnostic and therefore may ultimately produce a successful, biologically based, theory of mind. Mind, as with organic processes, is dependent on evolution, which is what makes a biological neuroscience solution radical.

The logic of interests in neuroscience

Leslie Brothers

Division of Psychiatry, UCLA-Sepulveda VA Medical Center, Sepulveda, CA 91343. **brothers@ucla.edu** **www.medsch.ucla.edu/som/npi/svapsych**

Abstract: Logical problems inherent in claims that biological neuroscience can ultimately explain mind are not anomalous: They result from underlying social interests. Neuroscientists are currently making a successful bid to fill a vacuum of authority created by the demise of Freudian theory in popular culture. The conflations described in the Gold & Stoljar target article are the result of alliances between certain apologist-philosophers, neuroscientists, and institutions, for the purpose of commanding authority and resources. Social analysis has a role to play in addressing logical issues in the philosophy of neuroscience.

By showing that the best contemporary neuroscience achieves only neural implementations of extant psychological narratives, not replacements of them, Gold & Stoljar (G&S) support their argument that the reach of the radical doctrine exceeds its grasp and, thus, that the doctrine is in effect a wager regarding the future.

Of course, one wants the reach of science to exceed its grasp. It is troubling, however, when the reach is portrayed *as though it were* the grasp. To use the title word of a well-known book, such a portrayal is "astonishing" in the same way that the results of sleight-of-hand are astonishing. A magician's sleight-of-hand is accomplished through the manipulation of attention: similarly, by narrowing their audience's attention to links between selected aspects of psychological theory and laboratory data, neuroscientist-communicators provoke admiring astonishment as they claim hegemony for biological neuroscience. The embeddedness of psychological concepts in the selection and interpretation of neural data, as in the Kandel example (see sect. 5.3.1), is overlooked.

I have shown elsewhere how the feat is accomplished in the case of emotion (Brothers 1997). The psychological concept, uncritically imported from lay culture, is woven into the laboratory observations from the beginning. Thus, it is no accident that neural results appear to validate – or, in the terminology of social theorists, to "naturalize" – the concept. In such a circular process, however, there is always something left over, some data that do not quite fit. The crucial question for the progress of inquiry is whether such data will be accounted for through creative extensions that do not undermine existing assumptions or used instead to challenge the underlying psychological concepts. In practice, cultural categories tend to be extremely resilient and accommodating. The resulting potential for self-confirming explanatory loops has serious implications for neuroscience, regardless of whether the radical or trivial neuron doctrine is operative.

The conflation of the doctrines and the occult mingling of psychological narratives with experimental data are logical problems.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

My conjecture is that their persistence is explained by underlying interests. The influence of social interests on the development of ideas has been explicated for other fields (Collins 1998; Restivo 1985). A similar analysis is overdue for neuroscience. On the one hand, it would be useful to study incentives and relations internal to the field (e.g., what factors influence the flow of resources between granting agencies, institutions, and individuals). Here, I suggest, a social context for the dearth of broad theoretical suggestions in neuroscience remarked upon by G&S (see sect. 6.7) can be found. On the other hand (but linked to the first consideration), there are relations between neuroscience as an extended institution and other groups, relations that involve status and access to resources. It is probably not a coincidence that recent claims for the hegemony of biological neuroscience in explaining the mind (buttressed, indeed, by the conflation of the neuron doctrines) coincide with a pressing popular demand for a new source of authority on precisely this subject. With the waning of Freudian authority, there appears to be a waxing appetite for neuroscientific authority. This is nowhere more apparent, to my observation, than in the community of psychotherapists, where neurobiological accounts of topics such as emotion and consciousness are in great demand. Some social scientists, concerned that they are "marginalized within the scientific community," incorporate neuroscience explanations into their accounts with the explicit aim of increasing their scientific legitimacy (Turner, in press). Agencies within the National Institutes of Health maintain their funding from congress at least in part by promoting biological neuroscience as the source of solutions to problems of human behavior, such as violence and addiction, in which society has an interest.

The incentives keeping neuroscience narratives tied to extant social narratives, and simultaneously dominant over them, are strong. Also, because the complexity, breadth, and increasingly technical methods of the field can be intimidating to nonbiologists, meaningful challenges to biological neuroscience's claims on the mind together with the hidden logical problems underlying them are unlikely to come from outside, unless from analyses such as G&S's. If neuroscience is not to stagnate, more such efforts, together with a comprehensive social analysis of the field, are needed.

Levels of description and conflated doctrines

John A. Bullinaria

Department of Psychology, University of Reading, Reading RG6 6AL, United Kingdom. **j.bullinaria@reading.ac.uk**
www.reading.ac.uk/AcaDepts/sx/Psych/PEOPLE/bullinaria.html

Abstract: It seems that I often say things that might mistakenly be thought to identify me as an adherent of the radical neuron doctrine. I take the opportunity to explain my position more clearly and argue that many apparent conflations of the radical and trivial neuron doctrines are merely the result of misunderstanding what is meant when neuroscientists talk about the relations between different levels of description. It follows that there may be considerably fewer followers of the radical doctrine than Gold & Stoljar suggest.

I agree with Gold & Stoljar (G&S) (sect. 4.2) that "if one is going to be a reductionist, one has to take the train of reduction all the way to the terminus of physics" and that this renders many fields, such as neurobiology, just local stops along the way. My position then follows from the facts that I am a self-confessed reductionist (see, e.g., Bullinaria 1986) and that I believe the "local stops" to be valid and useful intermediate levels of description.

From this viewpoint, psychology and linguistics are to neuroscience what chemistry and biology are to physics, or what the standard SU(3) × SU(2) × U(1) model of physics is to superstring theory or M theory (Duff 1998). (Incidentally, many theoretical physicists will disagree with G&S's assertion – in sect. 4.1 – that "physical theories rarely achieve [unification by] dissolution," and an outdated and irrelevant quote from Feynman does not inspire confidence in their familiarity with this area.) Generally, reductionists will take it for granted that each level of description follows from the more fundamental theory, but adjacent levels may look very different or be based on different concepts, and getting explicitly from one to the other is frequently a nontrivial task. Whereas all levels of description are important, and some are undoubtedly more fundamental than others, it is simply a matter of usefulness that decides which one uses under particular circumstances. Specifying linguistic rules in terms of the firing patterns of particular sets of neurons in the brain is no more useful than specifying the chemical processes necessary for manufacturing a new drug in terms of superstring wave functions.

Our choice of level of description is usually one of simplicity. If one can describe a system (e.g., a gas) adequately in terms of a small number of variables (e.g., pressure, temperature, volume) with simple relations, then why should we want to describe it in a more complicated manner (e.g., in terms of the motions of individual molecules), even if doing so is possible (either in principle or in practice). In section 2.3, G&S discuss the similar example of describing earthquakes. Given the limited working memory and processing abilities of the human brain (even when assisted by modern computational devices), a simplified account is often the only way we can hope to understand what is happening.

Naturally, the points of contact between the levels can be essential for formulating the correct descriptions at each level. If the relation cannot be made to work, one or both levels will need modification, or even discarding. This is natural scientific progress rather than something radical. In some fields the relations between levels remain crucial research areas, particularly at the most fundamental levels, where direct experimental validation is virtually impossible (see, e.g., Bullinaria 1986; Duff 1998). Although there will always be "in principle" relations between levels, the "in practice" relations are often simply not useful, owing either to computational limitations or to the enormity of the articulation of the relationship.

In each case, however, one still has to ask: "What would be gained or lost by analyzing the problem at a more, or less, fundamental level?" Psychologists are increasingly using neural network modelling techniques to explore human performance on numerous psychological tasks (such as language processing), and Newtonian dynamics might help pool players understand their skills, but worrying about general relativistic or superstring effects would only hinder both, despite those effects being universally present. The simplification is inevitably at the cost of approximation, and we must always ask if the approximation is good enough for its application, or if corrections from the more fundamental levels are necessary. In the examples given above, the approximations are clearly valid. On the other hand, for example, quantum processes are rarely described well by classical approximations. It is judging old approximations to be inappropriate that often leads to progress, for example, general relativity replacing Newtonian gravity, or connectionist neuropsychological replacements for old-style box-and-arrow cognitive models. The usefulness of each level depends on both the adequacy of the approximation and the degree of simplification.

There is nothing particularly new or controversial in the above, but it leads us on to a potential conflation of radical and trivial doctrines at each major level of description – psychology, neuroscience, biology, chemistry, physics – and also within each of these levels, for example, within physics – the standard SU(3) × SU(2) × U(1) model, supergravity, superstring theory, M theory. The conflation is not unique to the neuron doctrine. At each level we have a trivial doctrine that is simply an acceptance of our levels of description and a radical doctrine (as defined in sect. 2.2.2, para. 2) that has us discard a given level once the more fundamental level is fully developed and the relation between the levels is understood. The alleged conflations arise when one fails to state explicitly whether a given level will remain useful after it has been

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

understood in terms of the more fundamental level. I think too much is being read into what are often imprecise, throwaway comments intended merely to place a given piece of work in the broader context discussed above. In my case at least, statements of belief in an (ambiguous) neuron doctrine are simply never intended as an attempt to say anything radical. A statement of belief that a relation between levels will (eventually) be found is not the same as saying that it will render the less fundamental level superfluous. In this way, G&S's examples of evidence of commitment to the radical doctrine (in sect. 2.2.3) seem to be no more than commitments to general reductionist principles (that G&S dub the trivial doctrine), and it becomes difficult to find anyone guilty of the conflations alleged in section 2.3.

The philosophical debate about doctrines is thereby reduced to a practical debate about the usefulness of particular levels of description. It remains an important and interesting debate, in neuroscience as in many other fields, but many researchers, such as myself, will require further, more convincing, arguments that the creation of a new branch of philosophy is warranted.

Two radical neuron doctrines

Alex Byrne[a] and David R. Hilbert[b]

[a]*Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA 02139;* [b]*Department of Philosophy, University of Illinois at Chicago, Chicago, IL 60607.* **abyrne@mit.edu hilbert@uic.edu**
web.mit.edu/philos/www/byrne.html
www.uic.edu/depts/phil/hilbert.html

Abstract: Two radical neuron doctrines must be distinguished, strong and weak. Gold & Stoljar direct much of their attack at the former, but the Churchlands hold only the latter. The weak radical neuron doctrine remains a serious possibility.

Gold & Stoljar (G&S) describe the radical neuron doctrine in a number of slightly different ways, and we think that this hides an important distinction. On the one hand, the radical neuron doctrine is supposed to have the consequence "that a successful theory of the mind will make no reference to anything like the concepts of linguistics or the psychological sciences as we currently understand them," so Chomskyan linguistics "is doomed from the beginning" (sect. 2.2.2, para. 2, 3).[1] (Note that "*a* successful theory" must be read as "*any* successful theory," or the inference will fail.) On the other hand, the radical neuron doctrine is said to be the claim "that emergent psychological properties can be explained by low-level neurobiological properties" (sect. 2.3, para. 3). It is clear from the context that this can be more faithfully rendered as: psychological phenomena can be explained in (solely) neurobiological terms. However, this formulation of the doctrine does not have the consequence just mentioned. To adapt an example from the Churchlands (Churchland & Churchland 1994), that chemical phenomena can be ("in principle") explained in quantum mechanical terms does not imply that *all* successful theories of chemistry will employ (solely) quantum mechanical concepts and not use (classical) chemical concepts. It *does* imply that there is *some* successful theory of chemistry that does not employ chemical concepts, namely, quantum mechanics, but it is perfectly consistent with this that classical chemistry is successful, which, of course, it is. Admittedly, if quantum mechanics does explain chemical phenomena, there is a temptation to say that classical chemistry has a "second-rate, or place-holder, status" (sect. 1, para. 4). However, this is a somewhat tendentious description, because a quantum mechanical explanation of chemical phenomena need not detract from the explanatory power of classical chemistry.

So two radical neuron doctrines must be distinguished. The weak version says that the mind – psychological phenomena – can be (wholly) explained by neurobiology. The strong version is the conjunction of the weak version and the claim that *only* neurobiology can explain the mind. The strong version has the "radical consequences" discussed by G&S; the weak version does not. It seems to us that in the target article G&S's radical neuron doctrine is intended preponderantly to be the strong version.[2] However, we think that the Churchlands, and probably some other neuroscientists quoted by G&S, support only the weak version.

Now, if neurobiology (wholly) explains the mind, then it would seem to follow that any other theory that explains the mind either is *explained by* neurobiology, or else *explains* neurobiology. Therefore, although the weak version of the radical neuron doctrine does not have "radical" consequences, it does have a pretty strong consequence: Any true theory of the mind is either explained by or explains neurobiology. Because the psychological sciences certainly do not explain neurobiology, it follows from the weak version that any true psychological theory is explained by neurobiology or, following the Churchlands' (1994) and G&S's (sect. 4.2, para. 1) use of "reduction," any true psychological theory reduces to neurobiology. Thus, if linguistics cannot be explained by neurobiology (i.e., if linguistics does not reduce to neurobiology), and if the weak radical neuron doctrine is true, then linguistics is false.

With this distinction between the two versions of the radical neuron doctrine in place, consider the reductive formulation of the argument from unification (sect. 4.2). As G&S present it, because of the tendency of science toward unification, psychology must reduce to neurobiology, and so *any* successful theory of the mind must be solely neurobiological. G&S reply that by the same logic psychology must reduce to physics and thus *any* successful theory of the mind must be solely a theory in physics. Fair enough, but G&S are taking the conclusion of the argument to be the strong version of the radical neuron doctrine.

In fact, as presented by the Churchlands, the conclusion of the argument from unification is only the weak version: *Some* successful theory of the mind must be solely neurobiological. Therefore, if psychology reduces to physics, *some* successful theory of the mind must be solely a theory in physics, so here G&S's reply is of no force. The Churchlands do not think that a (smoothly) reduced theory is thereby rendered explanatorily defective, so a reduction of neurobiology to physics (or a reduction of the psychological sciences to neurobiology) would not threaten the power of neurobiology (or the psychological sciences) to explain the mind. For example, Patricia Churchland emphasizes that she "does not mean that there is something disreputable, unscientific or otherwise unsavory about high level descriptions or capacities per se" (1997, p. 128).

Thus the argument from unification in its reductionist formulation, taken as an argument for the weak radical neuron doctrine, is not as easily rebutted as G&S claim. We are not persuaded by it, largely because we think the notion of reduction (or explanation) is badly in need of clarification. On the other hand, neither have G&S persuaded us that the weak doctrine is false. Finally, we would like to endorse G&S's appeal for more philosophy of neuroscience, an admirable example of which is the target article.

NOTES

1. What do G&S mean by "successful"? Do they mean true, or rather something like very useful, widely accepted, and so on? We are unsure; for example, G&S's first paragraph suggests the latter interpretation, but section 1.2 suggests the former. In this commentary we have adopted the former interpretation, but our point can be made either way.

2. It is clear from the relation between the trivial and the radical doctrines that the second sentence of sect. 2.2.1, which concerns the trivial doctrine, can be transformed into the following claim about the radical doctrine: "According to [the radical] doctrine, to the extent that psychological phenomena will be explained at all, the science that will do so is [biological] neuroscience." This is the strong version: Only neurobiology can explain the mind.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

Why biological neuroscience cannot replace psychology

Nick Chater

Department of Psychology, University of Warwick, Coventry, CV4 7AL United Kingdom. **nick.chater@warwick.ac.uk**

Abstract: Gold & Stoljar argue persuasively that there is presently not a good case for the "radical neuron doctrine." There are strong reasons to believe that this doctrine is false. An analogy between psychology and economics strongly throws the radical neuron doctrine into doubt.

Gold & Stoljar (G&S) have provided an excellent case for believing that current arguments for the "radical" version of the neuron doctrine – that a successful theory of the mind will consist purely of biological neuroscience – are not persuasive. In this commentary, I take the next step and argue directly that the radical neuron doctrine is false.

In cognitive science it is standard to view the mind as a computational device. An analogy with conventional digital computers then immediately suggests that an equivalent of the radical neuron doctrine – what one might term the "radical transistor doctrine" – is patently absurd. The entire subject matter of computer science testifies that a successful theory of digital computation is not couched in terms of transistors or, indeed, in terms of electrical engineering at all. Instead, there is discussion of programming languages, computer programs, data structures, and so on, none of which has any interpretation at the level of transistors. However, this argument is not likely to persuade advocates of versions of the radical neuron doctrine, because they simply reject the analogy between the brain and digital computation.

Let us therefore consider instead a different domain: economics. Economists talk about notions such as "money," "price," "inflation," and so on. The laws of economics are defined over these and similar notions. Let us call the "radical physics doctrine" the view that economic phenomena should really be explained purely in terms of physical properties of the world. The deep problem with this doctrine is that economic notions do not have any physicalistic specification. Crudely, there is nothing *physical* about the note in my pocket that makes it worth five pounds sterling. After all, an atom-by-atom replica of the note created by some devious forger will be worth nothing, though it will pass successfully into circulation unless its origins are revealed. Moreover, if the Bank of England decides to mint a five pound coin, and remove the note from circulation, my note becomes worthless, even though its physical properties are unchanged. Similarly, and perhaps more strikingly, consider that the only physical correlate of my bank balance may be a complex distributed physical pattern on the hard disk of my bank's computer. The economic properties of my bank balance are (I hope) unchanged if my bank changes its computer system, or switches to a different kind of storage medium, but the physical correlate of my bank balance changes radically. Quite generally, economic concepts and laws can simply not be expressed at all in the language of physics; it would therefore seem to be a great mistake to attempt to explain economic phenomena in purely physical terms. The radical physics doctrine would appear doomed.

What is the upshot of the analogy? It places advocates of the radical neuron doctrine on the horns of the dilemma: They must either accept the analogy between the two doctrines and somehow find a way to accept the radical physics doctrine with respect to economics or find a disanalogy between this doctrine and the radical neuron doctrine, which shows the latter to be more acceptable.

The first option seems extremely unappealing. By staying with physical descriptions we are simply unable to talk about "money," "price," and so on, yet these notions appear to be central not just to present theoretical accounts of economics but to the very subject matter of the discipline. To put the point simply, we have no way of conceiving how we might predict and explain phenomena in terms of, say, elementary particles that we currently predict and explain in terms of supply and demand.

However, the second option is also difficult to defend. There seems to be no more reason to suppose that psychological notions, such as "memory," "attention," and so on will one day be replaced by rigorous talk about neurons than there is to suppose that economic notions, such as "money," will be replaced by talk about physics.

One attempt to make the overthrow of psychology by neuroscience seem plausible points to the provisional, partial, and generally unsatisfactory state of psychological theory, but the very same criticism can be levelled at economics. In both cases, what seems to be required is more and better theory at the same level of analysis; it seems patently self-defeating to attempt a radical shift to a different and more basic level of analysis.

Another attempt to make the overthrow of psychology by neuroscience seem plausible points to the fact that neuroscientists are gradually clarifying how psychological notions such as "memory" have a neural basis (in terms, for example, of long-term potentiation; Bliss & Lømo 1973). This is of no help, however, because this kind of knowledge is already in place in the economic case – we already know that the "physical basis" of my being able to buy a newspaper consists of the possession of coins or notes with particular physical properties. The problem is that in neither the psychological nor the economic context are the physical properties appropriate for couching relevant generalizations: The physical properties of a note do not make it worth five pounds, and the physical properties of a memory do not make it a memory of visiting London or buying a filing cabinet. In sum, the radical physics and the radical neuron doctrine seem equally unattractive as guiding principles for scientific research; in both cases, accepting the doctrines immediately undercuts the theorist's ability even to talk about the phenomena of interest, let alone to explain them.

G&S end their target article with a plea for further work in what they see as the underdeveloped field of the philosophy of neuroscience, to be concerned with the "presuppositions and philosophical problems of neuroscience itself" (sect. 6.2, para. 2). I suspect that the philosophy of neuroscience appears underdeveloped because there simply is nothing to develop. Once the confusions and ambiguities concerning the relationship between neuroscience and other perspectives on the mind are clarified and are cleared away, as G&S have so ably done, the link between neuroscience and philosophically interesting issues is broken. Perhaps a putative philosophy of neuroscience would be no more substantial than a putative philosophy of cellular processes in the lung or the heart.

How trivial is the "trivial neuron doctrine"?

Steven G. Daniel

Department of Philosophy, University of Nevada, Reno, Reno, NV 89577-0056. **daniel@unr.edu**

Abstract: I argue that Gold & Stoljar's "trivial neuron doctrine" is not in fact trivial. Many familiar positions in the philosophy of mind run afoul of it, and it is unclear that even those whom Gold & Stoljar identify as adherents of the trivial neuron doctrine can be comfortably described as such.

Gold & Stoljar (G&S) distinguish between a radical and a trivial version of the "neuron doctrine" and argue that the former is implausible. This, I think, is right. At the same time, I wonder whether even the trivial neuron doctrine is all that trivial.

According to G&S, the trivial neuron doctrine holds that the correct theory of the mind "will turn out to involve any one of a very large number of possible combinations of scientific concepts" (sect. 2.2.1, para. 1). This seems straightforward, yet, although the scientific concepts in question can come from any one of a num-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

ber of disciplines, these must be disciplines falling within the purview of cognitive neuroscience. Hence, as an example of a view that is incompatible with even the *trivial* neuron doctrine, G&S cite "a certain version of the *artificial intelligence program,* according to which both neurobiology and the details of psychology are in principle irrelevant to the construction of theories of mentality in the most abstract sense" (sect. 2.2.1, para. 6). G&S do not describe more precisely the kind of view they have in mind, but we can consider a few of the possibilities.

Perhaps the artificial intelligence program at issue assumes that a system's ability to pass an unrestricted Turing test is at least a sufficient condition of its being intelligent. Such a *behavioral* criterion of intelligence tends away from details about inner processes and mechanisms. Furthermore, Paul Churchland (1996) has argued that the Turing test does not constitute a satisfactory criterion of intelligence, that a system's ability to pass the Turing test is neither a necessary nor a sufficient condition of its being intelligent. He notes, "Whatever [the Turing test's] merits or demerits as a criterion of intelligence, it is independently clear that we are forced to fall back on 'behavioral similarity to a paradigm case' only so long as we lack an adequate theory of the paradigm case, an adequate theory of what intelligence is and how it is realized in physical systems" (pp. 234–35). The paradigm case Churchland has in mind is obviously the human case, the theory a neuroscientific theory. However, there is no reason brashly to assume that our best neuroscientific theory of the paradigm case of human intelligence will be particularly useful at explaining intelligence in general, and the conclusion that the ability to pass an unrestricted Turing test is not even sufficient for intelligence is controversial, not trivial.

It seems likely that certain varieties of functionalism in the philosophy of mind (such as "analytical functionalism") also run afoul of the trivial neuron doctrine. If so, this again calls its triviality into question.

Now, G&S do say that they "are not claiming that the trivial neuron doctrine is trivial in the logical sense" (n. 15), but I wonder whether it is rightly described as trivial in any sense. Perhaps it is supposed to be trivial only in this sense: Ask any researcher in cognitive neuroscience whether the trivial neuron doctrine is true, and they will answer in the affirmative. Perhaps they would. Nevertheless, it is not obvious that all of those writers whom G&S describe as proponents of (any version of) the neuron doctrine can be, in the end, comfortably described as such.

Returning to the topic of artificial intelligence, Patricia Churchland (1986) has remarked that "if human brains and electronic brains both enjoy a certain type of cognitive organization, we may get two distinct, domain-relative reductions. . . . In and of itself, the mere fact that there are differences in hardware has no implications whatever for whether the psychology of humans will eventually be explained in neuroscientific terms" (p. 357). Churchland here allows that human intelligence might be explained neuroscientifically, but she does not say that the concepts of neuroscience will have to figure prominently in explaining all forms of artificial intelligence. (Nor should she; a concept's playing a significant role in a theory about the intelligence of nonbiological systems would be prima facie evidence of its independence from neuroscience.) In Churchland's view, there *will* always be a kind of psychophysical reduction: Mental property M will reduce to (for example) a particular neurobiological property, P, in humans and to some other physical property in artificial domains. Of course, the integrity of the notion of domain-relative reduction is doubtful (see Blackburn 1991). Still, what seems important to Churchland is that relatively high-level mental properties reduce to lower level properties in every domain. The lower level properties do not have to be neurobiological properties, though.

Is this consistent even with the trivial neuron doctrine? If it is, then why call it a "neuron" doctrine at all?

More to the point, why can we not be pluralists? Why can we not embrace the artificial intelligence program, functionalist theories of the mind, and also (a suitably liberalized version of) the trivial neuron doctrine so long as we allow that these theoretical approaches address a wide variety of issues at different levels of abstraction? The proponent of the trivial neuron doctrine might argue that embracing (for example) the artificial intelligence program and functionalism would lead us away from naturalism, the view that "to the extent that we will be able to understand the world, it will be empirical science (and not, say, religion or philosophy) that provides that understanding" (G&S, sect. 1, para. 3)[1] and naturalism is built into the trivial neuron doctrine. However, if the trivial neuron doctrine's naturalism refuses any role at all to these more abstract forms of theorizing about the mind, its slant on naturalism is anything but trivial.

NOTES

1. For an alternative construal of naturalism, see Papineau (1993).

Reductionism and the neuron doctrine: A metaphysical fix of Gold & Stoljar's trivial–radical distinction

James Fahey and Michael Zenzen

Department of Philosophy, Psychology and Cognitive Science, Rensselaer Polytechnical Institute (RPI), Troy, NY 12180-3590.
{faheyj2;zenzen}@rpi.edu

Abstract: The trivial neuron doctrine (TND) holds that psychology merely *depends on* neurobiology. The radical neuron doctrine (RND) goes further and claims that psychology is superfluous in that neuroscience can "replace it." Popular among RND notions of "replacement" is "reduction," and in our commentary we challenge Gold & Stoljar (G&S) to make clear their distinction between merely *depends on* (TND) and *is reducible to* (RND). G&S give us a TND–RND distinction that is a *distinction without a difference;* a defensible TND–RND distinction must have a metaphysical basis. We suggest a denial of *compositionalism* as such a basis.

We look at two interrelated questions: (A) Have Gold & Stoljar (G&S) hit on a common equivocation in the philosophy of neuroscience, or is the trivial neuron doctrine (TND)–radical neuron doctrine (RND) distinction a *distinction without a difference* and (B) do G&S's arguments against RND hold up? Our answers to (A) are "yes" and "yes." G&S's intuition that close scrutiny of the neuron doctrine reveals that *a kind of* TND–RND distinction *is* well founded. We argue, however, that G&S's account of this distinction suffers from lack of metaphysical grounding. Our vehicle for answering (A) is our denial of (B). G&S examine versions of one prominent family of arguments in favor of RND, "arguments from the unity of science," and find each to be uncompelling. We look at their specific denial of unification as reduction (sects. 4 and 4.2) and, by undercutting their argument, put their TND–RND distinction in jeopardy. We then show how both their denial and a version of the TND–RND distinction can be saved, if their argument is recast to allow a denial of compositionalism, the view that the properties of things are completely derivative of the properties of their respective constituent parts.

To summarize G&S's argument from the unity/reduction of science (U/R) in support of RND: (1) The tendency toward unity as reduction has led to scientific theories with greater explanatory power. (2) This welcome "global tendency" will be brought to fruition most expeditiously through an accretion of more local reductions, and in the sciences of mind this tendency would be maximally supported by a successful reduction of the psychological sciences to neurobiology. Therefore, (3) a successful theory of mind will be solely neurobiological.

In their denial of U/R, G&S focus on (2). They point out that adoption of TND does not require that psychology be reducible to neurobiology. Indeed, proponents of TND often explicitly deny that such reduction is possible. RND, however, does require reduction; it requires that all substantive psychological explanations be reformulable as neurobiological explanations. G&S argue,

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

however, that the radical's acceptance of this kind of intertheoretic reduction subverts his or her own view on two counts: (1) If the "unity" of science results from the fact of a thoroughgoing intertheoretic reduction, then there is only one genuine, basic science – (micro)physics. (2) Moreover, supposing that such microphysical reduction is true, this by itself reveals nothing about reductive relations among nonbasic sciences such as psychology and neurobiology.

Is the kind of reduction G&S refer to in U/R and its denial adequate to the task at hand? We think some holders of RND might respond as follows:

> Yes, the psychological is reducible to the neurobiological, but what is at issue is not the reductive translation of one theory of explanation into another but rather the claim that the psychological is nothing but the neurobiological. Explanations are essentially intentional; what counts as a good one depends on our interests, on the projects that engage us. Given this, translations may fail because they fail "to satisfy," so we can agree with those who hold to the TND that in this sense psychology merely depends on neurobiology and is not reducible to it. However, whether or not reduction is the case is independent of our "satisfactions." Rather it depends on whether the intrinsic properties exemplified by such things as minds(-brains), derive from or are inherited from the neurons, and so on, that constitute them.

We believe that this is a formidable argument. Moreover, if it holds, G&S's claim that there is a TND–RND distinction is weakened substantially and their argument against U/R is undercut. If holders of both TND and RND can deny that psychology is intertheoretically reducible to neurobiology, then, in this respect at least, the TND–RND distinction is a distinction without a difference.

Now suppose we recast both G&S's statement of U/R and subsequent denial in terms of an account of reduction such as that given above. Our new argument from unity as metaphysical reduction (U/MR) proceeds as before, except now we understand that science tends toward unity because it continues to uncover the compositional nature of things.[1] Can this reformulated U/MR support a genuine TND–RND distinction?

Given that G&S allow that proponents of both TND and RND hold that "mental phenomena are identical to neural phenomena" (sect. 3.2), it seems we are no better off. Again it seems that each holds that psychological properties not only depend on but also derive from neurobiological ones, and again the distinction evaporates.

To save the TND–RND distinction, we must pay a metaphysical price. If holders of TND adopt the view that, while the psychological depends on the neurobiological it does not derive from it, then we do have a sound, metaphysical basis for the TND–RND distinction. However, this commits such nonreductive (physicalist) proponents of TND to noncompositionalism (emergentism), and many find this unappealing (Kim 1992). Nevertheless, it seems that our option is forced: Either pay an appropriate metaphysical price or give up the TND–RND distinction.

One advantage of accepting a noncompositionalist outlook is that a reformulation of G&S's rebuttal to U/MR looks more promising. G&S question (1) the cogency of RND as providing a "local-reductive-stop" and (2) the claim that psychology derives from neurobiology and because of this is more basic. As regards (2), G&S claim that this is an empirical question, and objectors to RND can justifiably argue that the jury is still out. Moreover, now (1) seems more powerful. If the properties of everything derive from the properties of the microphysical, then neurons are no big deal and RND is rendered otiose.

NOTE

1. We agree with G&S (n. 30) that "the fact that As are made up of Bs does not entail that As are to be explained in terms of Bs" and add that mere constitution does not guarantee inheritance of properties in a systematic way. However, further clarification of this point requires a discussion of natural kinds that cannot be undertaken here.

Of skyhooks and the coevolution of scientific disciplines

Donald R. Franceschetti

Department of Physics and Institute for Intelligent Systems, The University of Memphis, Memphis TN 38152. **dfrncsch@memphis.edu**

Abstract: The history of the natural sciences repeatedly shows that the unification of a higher level theory with a lower level theory by reduction does not eliminate the need for the higher level theory nor preclude its further development, leading to changes in the understanding of the lower level. The radical neuron doctrine proposes that the future science of psychology or linguistics will derive principally from the evolution of understanding at the neural level and not from current theories based on the observation of behavior. It is far more likely that the two bodies of theory will coevolve in semiautonomous fashion.

The notion of unification of the sciences through the reduction of all phenomena to the dynamics of a small set of material particles has played a role in scientific thought since the time of Plato's *Timaeus* or somewhat before. As was noted by Mauldin (1996), this unification is now considered the single goal of physics by many of that discipline's practitioners. It is certainly prominent in the development of the molecular basis for biology as well. From a more cynical viewpoint it might be considered a defining myth, being taught to subsequent generations of young scientists and being trotted out for the edification of funding agencies and the general public, whenever support for an area of research appears to be threatened.

As Simon (1996) has noted, however, progress in science occurs far more often from the human scale downward than from the bottom up, that is, by "skyhooks" rather than by building "skyscrapers." That this is so is a reflection of the relative autonomy of the different levels of description, and it is a good thing, because increases in understanding at the most elementary levels are quite hard to come by. Thus our understanding of, say, the chemical bond between atoms is not subject to profound change each time a new subatomic particle is discovered. If the history of the physical sciences can be taken as prototypical, there is little reason to expect a reconceptualization of the higher level to result from even dramatic progress at the lower.

Two examples illustrate the point well. In the mid-nineteenth century, English physicist J. C. Maxwell summarized a century of experimentation with electrical and magnetic phenomena in a set of four equations that predicted the existence of electromagnetic waves that would propagate through empty space with precisely the velocity of light. Acceptance of the electromagnetic nature of light was rapid and total, to the extent that any advanced textbook of optics now begins with Maxwell's famous equations. Nonetheless, the evolution of optical concepts has a very long history, ranging from the law of reflection known to Euclid to Snell's law of refraction and Fermat's principle of least time. Any course in optics is largely conducted with the use of optical concepts that antedate the current picture of the "true nature" of light, whereas the principal impact of the reduction or unification is to justify the optical principles that were discovered empirically. Likewise the reduction of chemistry to atomic physics is thoroughly accepted, but this does not eliminate the value of a plethora of chemical concepts to the practicing chemist. The phlogiston theory was replaced not because of progress at the atomic level but because Lavoisier's identification of oxygen as an element explained a much broader range of phenomena. On the biological side, lessons learned from inheritance in pea pods and fruit flies strongly guided the search for the chemical nature of the gene and still provide the basis for genetic counseling.

From a more fundamental standpoint, the case can be made for the coevolution of scientific theories, with the higher level often driving the agenda of research at the lower. This is particularly true in the case of emergent properties, such as phase transitions, melting, boiling, and superconduction, which are almost invariably

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

first noted in observational data at the higher level and then assiduously sought at the lower. The controversial proposition within the radical neuron doctrine might then be the issue of whether the behavioral-level theory, that is, psychology or linguistics, will evolve continuously from the current psychology or whether a Kuhnian revolution will be needed to obtain theories that work better than the current ones.

In addition to the argument from precedent that a science of psychology or of linguistics will still be needed after neuroscience reaches its maturity is the further fact that an autonomous formal theory of behavior is possible. Thus the theory of computation for Turing machines can be developed without reference to any particular substrate that embodies the Turing machine, and the existing theory of neural networks can be developed for artificial neurons built out of a variety of materials. The intriguing possibility (Penrose 1989; 1994; Penrose et al. 1997) [See also *BBS* multiple book review of Penrose's *The Emperor's New Mind BBS* 13(4) 1990] does exist, however, that neural processes might include ones that are noncomputational in nature, that is, not simulatable to arbitrary accuracy by a Turing machine as artificial neural networks are. If this is true, then, because all current microphysical theories are computational, what will be required is a coevolution, or more likely corevolution, of several levels of scientific theory, possibly driven by advances in the psychology of the conscious mind.

What neuron doctrines might never explain

Keith Gunderson

Department of Philosophy, University of Minnesota, Minneapolis, MN 55455.
gunde002@maroon.tc.umn.edu

Abstract: My focus is on the inability of neuron doctrines to provide an explanatory context for aspects of consciousness that give rise to the mind–body and other minds problem(s). Neuroscience and related psychological sciences may be viewed as richly contributing to our taxonomic understanding of the mind and conditions underlying consciousness, without illuminating mind–body and other minds perplexities.

In describing the neuron doctrine Gold & Stoljar (G&S) rightly refrain from making it a sitting duck: "The idea is not of course that neuroscience will explain everything about the mind; perhaps there are aspects of the mind we will never explain" (sect. 1, para. 1).

This is perhaps a harmless disclaimer, and certainly not the focus of their fascinating article, yet I find it irresistible to ask whether anything useful can be said about those aspects of the mind one might be considering when one thinks of that perhaps inexplicable residue? Is it related in some way to the G&S charge of hyperbole with respect to Paul Churchland's "optimism about current neural science" and G&S's remark that "it is only a slight exaggeration to say that we are almost completely ignorant about how the brain produces mental life" (sect. 1.3, para. 2)? Some of this must have to do with the dreaded C word – *consciousness* – and the attendant issues of the mind–body (MB) problem and companion perplexities concerning other minds (OM), topics often conspicuously absent from otherwise global portrayals of the mind. After all, the wish to avoid the wild metaphysical conclusions that these puzzling aspects of mind might seem to foster has motivated philosophers and their various scientific allies to argue for a naturalistic and materialistic metaphysics in the first place, which is viewed as lending support to the neuron doctrine. So too the neuron doctrine can in turn be viewed as a way of providing tough-minded, detailed support for a naturalistic and materialistic metaphysics. In any case such metaphysicians and scientists wish to avoid, as G&S cite Higgenbotham as claiming, a Cartesian research assumption that the mental is a separable substance from the physical (sect. 2.2.1, para. 4). This goes without saying, but where does that leave us? I would like to hear what G&S have to say about (1) whether anything in neuroscience broadly construed sheds light on MB or OM problems and (2) the relation of this (here grossly underdescribed) conundrum to our ignorance concerning how the brain produces mental life.

There appear to be two sets of issues at stake, which in their own way tend to be as conflated as those features of the neuron doctrine (the trivial and the radical) that G&S usefully clarify for us. First, there are those issues involving perspective or points-of-view problems, problems of how a first-person ontology (ours, bats, etc.) involving consciousness might be explained within a third-person framework, problems of objectifying subjectivity as Thomas Nagel, John Searle, and others have described it. If this is not an issue, why does this not matter or lend credence to a scary antiphysicalism or nonnaturalism? These can be collectively described as "investigational asymmetries problems."

How do we explain the radical intuitive differences in the felt texture of conscious experiences from anything we can imagine being churned up in neuroscience? It seems that what we can expect from neuroscience and related disciplines are either accounts of the alleged conditions *underlying* (not obviously identical with) conscious experience or accounts of how the results of those experiences diachronically occur and can be usefully characterized. Both of these may be viewed as problems of characterization or taxonomy. Current interdisciplinary projects addressing them have considerable power and charisma and are obviously helping to decode fundamental features of mentality. For this compelling reason many believe that as such projects develop they will explain whatever there is to explain about the mental save for its underlying physics. However, perhaps all this illumination could occur without a ray of light being shed on the investigational asymmetries problems and MB and OM problems that are their fallout. Locke noticed this division of issues in his own way in the seventeenth century. At the beginning of his *Essay Concerning Understanding*, he wrote that he was not going to "meddle with physical considerations of the mind." What did he mean by that? He meant that he was going to produce a developmental taxonomy or characterization of the mind that would explain knowledge acquisition sans any need for innate ideas or principles. The result was his elaborate parsing of the human mind in terms of simple and complex ideas *without* addressing the MB problem. Why? It seems pretty clear that it was because he appreciated the pickle Descartes got into when he did address that problem and tried to explain mind–body interaction, the connection between consciousness and the brain or body via his own precursor of a neuron doctrine. Descartes located the soul or consciousness in the pineal gland and claimed that it affected both involuntary and voluntary bodily movements via the "animal spirits," forerunners of modern neural firings. In a way, in spite of his metaphysical dualism, Descartes tried to physiologize the soul. However a convincing account of that proposed connection never panned out as his young disciple/critic Princess Elizabeth forcefully pointed out in her lively correspondence with him. The "point of contact" problem, as we might call it, was never solved and seemed forever unsolvable within Descartes' research program.

Are current non-Cartesian research programs any more convincing in illustrating how first-person points of view can be anchored in physicalistic third-person waters? Might our allegedly sweeping ignorance concerning how the brain produces mental life mentioned by G&S at the outset be indicative that they are not? Do they solve any contemporary version of the "point-of-contact" problem? There are versions of it that I have no space to delineate here, but perhaps we should forego that metaphor and any others like it.

Aspects of first-person points of view wherein the primary ontology of mind resides and badgers our comprehension cannot be accounted for in any kind of current physicalist reduction model (macro- to micro- or any other kind). Could this be because points of view are neither macro- nor micro-anything yet not thereby nonnaturalistic or dualistic in nature? If this is so, it must mean that

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

one has to reformulate in a radical way the structure of MB and OM problems and do so in such a way that, surprisingly, neuroscience as we now know or imagine it is utterly irrelevant to them.

The neuron doctrine is an insult to neurons

Stuart Hameroff

Departments of Anesthesiology and Psychology, The University of Arizona, Tucson, AZ 85724. **hameroff@u.arizona.edu**
www.u.arizona.edu/~hameroff

Abstract: As presently implemented, the neuron doctrine (ND) portrays the brain's neurons and chemical synapses as fundamental components in a computer-like switching circuit, supporting a view of brain = mind = computer. However, close examination reveals individual neurons to be far more complex than simple switches, with enormous capacity for intracellular information processing (e.g., in the internal cytoskeleton). Other poorly appreciated factors (gap junctions, apparent randomness, dendritic-dendritic processing, possible quantum computation, the living state) also suggest that the ND grossly oversimplifies neuronal functions. In the quest to understand consciousness, the presently implemented ND may throw out the baby with the bath water.

Whether a successful theory of mental phenomena will be solely neuroscientific (the "radical neuron doctrine") or will require additional psychological features is a moot question. In either case the neuron doctrine (ND) currently in vogue, and as presently foreseen, may be too watered-down to explain mental phenomena. Neuroscience is not being applied deeply enough.

Proponents of what Gold & Stoljar (G&S) describe as the radical ND (e.g., the Churchlands: P. M. Churchland 1995; P. S. Churchland 1986) consider only certain activity at neuronal surfaces, ignoring internal features, other details, and factors related to neurons as living cells. In the ND picture, neuronal axon membranes "fire" and propagate traveling action potential "spikes" on the axonal surface. Upon reaching presynaptic axon terminals, spikes cause release of neurotransmitters into the synaptic cleft. These in turn trigger dendritic membrane events in a second neuron, which can culminate in another axon firing, more spikes, further neurotransmitter release, and so on. Networks of synaptically connected neurons self-organize and, by adjusting synaptic strengths ("synaptic plasticity"), can account for learning.

The ND view conveniently lends itself to artificial neural networks: Bit states in silicon are analogous to synaptic transmissions or axonal firings. Indeed, neural networks in the brain have inspired a generation of parallel distributed "neural" networks in computers (Churchland & Sejnowski 1992). This connection (brain = mind = computer) seems implicit in the ND.

This is fine, except that the ND brain = mind = computer view still leaves us "almost completely ignorant about how the brain produces mental life" (sect. 1.3, para. 2). The ND expectation is that consciousness emerges at some critical level of computational complexity, and that emergence itself, rather than any property specific to neurons, accounts for conscious experience. In this view neurons are equivalent to switches and consciousness is destined to emerge in silicon computers. However, there are no testable predictions for such emergence. We can only wait for it to happen. The neuron doctrine is a bluff.

Perhaps we should look more closely at neurons. The present-day ND characterization is a cartoon, a skin-deep portrayal that simulates a real neuron much as an inflatable doll simulates a real person. Here are five features of neurons ignored as "messy details" in the ND. Are they necessary to explain consciousness? We do not yet know.

1. Neurotransmitter vesicle release is seemingly probabilistic. In many neurons only about 15% of axonal action potential spikes reaching presynaptic terminals cause actual release of neurotransmitter vesicle. (Apparent randomness exists at all scales in the nervous system; see Arieli et al. 1996; Barinaga 1996). Although probabilistic randomness per se is not a problem for the ND, variability could actually reflect deeper levels of organization such as patterns of spikes (rather than just average frequency), cytoskeletal states, or quantum indeterminacy in the vesicle release mechanism (Beck & Eccles 1992).

2. Apart from chemical synapses, primitive electrotonic gap junctions may be important. Gap junctions are portholes between cells, windows of cytoplasm joining neurons into one synchronously firing "giant neuron" (Kandel et al. 1991). Gap junctions are important in embryological development but are considered background players to more abundant chemical neurotransmitter synapses in developed brains. However, gap junctions remain functional and prevalent throughout all mammalian brain areas (Micevych & Abelson 1991), and their true importance may be hidden. Ironically, gap junctions connect neurons in a kind of "reticulum," or "syncytium," which Ramón y Cajal discarded in favor of individual, discrete neurons. However, gap junctions are transient; they come and go, regulated by cytoskeletal structures.

3. Dendritic-dendritic processing may be essential for consciousness, with axonal firings supporting automatic, nonconscious activities (see, e.g., Pribram 1991). Eccles (1992) portrayed "dendrons" (units of 100 pyramidal cell apical dendrites) as the functional units of consciousness. Unique organelles (dendritic lamellar bodies; DLBs) are found only on opposite sides of each dendritic-dendritic gap junction, anchored to cytoskeleton (De Zeeuw et al. 1995).

4. Any possible significant role for glial cells (80% of the brain) is cast aside in the ND. (Glia connect with neurons by gap junctions.)

5. The ND treats each neuron as a "black box," ignoring internal activities. Dynamic structural organization, including synaptic plasticity, is a function of the cytoskeleton. Microtubules and other cytoskeletal structures determine neuronal architecture and synapses; service ion channels, synaptic receptors, and gap junctions; transport and release neurotransmitter vesicles; convey intracellular signaling; and regulate gene expression (Fig. 1). The synaptic plasticity essential to Kandel's learning – as well as long-term potentiation (LTP) in mammalian neurons, etc. – depends on dynamic activities of the cytoskeleton. Signaling and communication occur through microtubules (see, e.g., Maniotis et al. 1997), and numerous theoretical models suggest that micro-

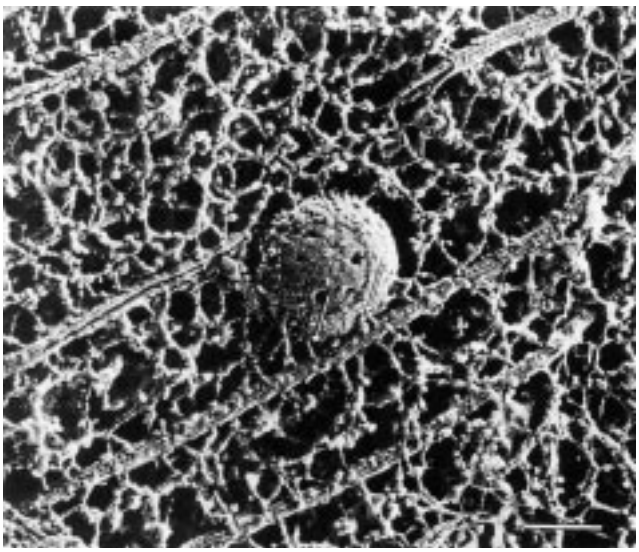


Figure 1 (Hameroff). Cytoplasm within a neuron showing approximately five microtubules (diagonal, lower left to upper right), with vesicular organelles being transported by motor proteins (arrows) attached to, and organized by, microtubules. Scale bar = 100 nmu. From Hirokawa (1991) with permission.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

tubules are well-designed information processors (see, e.g., Rasmussen et al. 1990).

Consider a single-cell *paramecium,* which swims gracefully, avoids predators, finds food, mates, and has sex, all without a single synapse. Remarking on the complex behavior of motile protozoa, C. S. Sherrington (1951) said: "Sense organs . . . seem to inspection wanting. Of nerve there is no trace. But the cell framework, the cyto-skeleton, might serve. There is therefore, for such mind as might be there, no need . . . to . . . say 'the apparatus for it is wanting.' " If the cytoskeleton can be so useful in protozoa, what might it be doing in massive parallel arrays within neurons? Are neurons stupid in comparison to protozoa?

Are there implications of neurons being alive? Could consciousness depend on some essential feature of life? For example, the unitary nature of living systems may involve quantum states in cytoplasm, and models of consciousness propose quantum computation in microtubules (Hameroff 1998a; 1998b; 1998c; Hameroff & Penrose 1996a; 1996b; Penrose & Hameroff 1995). Quantum computation may soon be an important computational paradigm, and comparisons of the brain/mind to a quantum computer seem inevitable.

Mere switch? The neuron is a whole universe.

The nontrivial doctrine of cognitive neuroscience

Valerie Gray Hardcastle

Department of Philosophy, Virginia Tech, Blacksburg, VA 24061-0126, and Department of Philosophy, University of Cincinnati, Cincinnati, OH 45221-0374. **valerie@vt.edu**

Abstract: Gold & Stoljar's "trivial" neuron doctrine is neither a truism in cognitive science nor trivial; it has serious consequences for the future direction of the mind/brain sciences. Not everyone would agree that these consequences are desirable. The authors' "radical" doctrine is not so radical; their division between cognitive neuroscience and neurobiology is largely artificial. Indeed, there is no sharp distinction between cognitive neuroscience and other areas of the brain sciences.

Gold & Stoljar (G&S) claim that philosophers and cognitive scientists alike have confused what they dub "the neuron doctrine" with two different claims: that, in the end, the mind sciences will become cognitive neuroscience and that, in the end, they will become neurobiology. They believe the former position is trivial but that the latter is not. G&S believe that the doctrine is trivial because cognitive neuroscience can include any concept from psychology or biology. Hence, any theory we already accept about the mind is already part of cognitive neuroscience. If we adopt cognitive neuroscience as the appropriate framework for explicating mindedness, then nothing serious need change in the way we are currently investigating human psychology. Implicit in their discussion of Kandel's sea slugs is the claim that even theories that prima facie appear to be a good case for noncognitive neurobiology reducing psychology are actually part of cognitive neuroscience, because these theories make essential use of psychological concepts. The doctrine is both ubiquitous and trivial.

I disagree on both counts. Though I do believe (and have argued extensively elsewhere [Hardcastle 1996]) that a brain-centered mind science is the best way to understand the mind, I do not concur that their "trivial" doctrine is at all the self-evident standard in cognitive science. In fact, it is quite controversial, and rightly so.

There is an honorable tradition in artificial intelligence (AI), linguistics, and psychology disavowing any connection to the brain. Regardless of whether one believes that this tradition is ultimately healthy, it has flourished for many decades now. I need not rehearse the reasons the "East Pole" philosophers and scientists give for concluding that they can do cognitive science perfectly well without the brain. Suffice it to say that they maintain that including brain data into cognitive investigations would only bury psychological theories under masses of irrelevant data, especially if one believes that minds are multiply instantiable in all sorts of nonbiological things.

I take it that G&S's response is that we can simply join these disciplines with what we know about the brain to create cognitive neuroscience. However, we can do so only with loss. First, we would lose the generality of the higher ordered theories; they would now be restricted to brained creatures. Second, and by my lights more importantly, we would constrain the paths that both neuroscience and psychology could follow. Many have argued that psychology and neuroscience are autonomous sciences, that each should develop its theories independently of what the other hypothesizes. Churchland's (1986) "reductive coevolution" is the dream of a few, but others see it as a nightmarish straitjacket.

These sentiments come from the nice division that G&S highlight between metaphysics and methodology. Even if, for us, minds are brains, that fact does not entail that theories of minds will be theories of brains. What things are made of does not determine how we explain these things, or else all we would need for scientific understanding would be collapsible Hilbert spaces. The bottom line is that it is just false that everyone would agree that the trivial neuron doctrine is true, or even desirable.

However, it is not the case that the radical version of the neuron doctrine need be so radical, either. G&S argue that Kandel's work on the neurophysiology of learning is part of cognitive neuroscience because he develops a neurobiological theory of learning, which is the province of psychology. Similarly, I suppose, theories of immunology should be considered psychological theories as well, because our immune systems learn to create antibodies specific to particular viruses. I doubt that philosophers should be the ones to decide which discipline gets to claim which terms as its own. Kandel explores the neurobiology of learning in sea slugs and then generalizes what he learns to other biological systems. Though what he has discovered might have implications for how we understand the distinction between long-term and short-term memory, *Aplysia* are not particularly cognitive creatures. The fact that a discussion of a neural ganglion might use terms like "content" or "memory" or "learning" or "communication" does not thereby make it psychological, for these words could refer to the "basic . . . functional properties of neurons, ensembles, or structures" (sect. 2.2.2, para. 3). This is not for us in the armchair to decide.

Though I agree that the philosophers and others who discuss the relation between psychology and neuroscience do not maintain a consistent picture of what they mean by "neuroscience," what is meant cannot be resolved by terminological fiat. What counts as cognitive neuroscience as opposed to behavioral neuroscience as opposed to neurophysiology is quite gray; each domain shades into the others. I would think that at best one should be indifferent about the category into which we place his research. At the least, no deep philosophical questions should be resolved by it.

Neuron doctrine: Trivial versus radical versus do not dichotomize

Barry Horwitz

Laboratory of Neurosciences, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892. **horwitz@helix.nih.gov**

Abstract: Gold & Stoljar argue that there are two (often confused) neuron doctrines, one trivial and the other radical, with only the latter having the consequence that non-neuroscientific sciences of the mind will be discarded. They also attempt to show that there is no evidence supporting the radical doctrine. It is argued here that their dichotomy is artificial and misrepresents modern approaches to understanding the neuroscientific correlates of cognition and behavior.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

The dichotomy offered by Gold & Stoljar (G&S) – trivial versus radical neuron doctrine – represents an artificial view of how neuroscience and cognitive behavior are (and will be) linked. Rather than thinking in terms of a dichotomy, more justice to the way this type of science is actually carried out can be provided by viewing the complex connections between neuroscience and psychology as part of a hierarchy of conceptual relationships. Each level of the hierarchy is critical, as are the theories that bridge adjacent levels.

Besides dichotomizing a continuous conceptual framework, a second major problem that clouds the G&S target article is the apparent relish with which they use a number of highly biased (at least to me) terms and definitions to formulate their position; for example, "trivial" versus "radical" neuron doctrine. Whether the terms are theirs or others', and even though philosophers have every right to make exacting definitions, G&S's terminology distracts one from assessing their arguments, making it difficult for those of us who might concur with some of what they say to want to agree with them.

Psychological explanation is concerned with the behavior of the organism as a whole; neurons (or even small neural networks) do not perceive, remember, use language, and so on. These terms describe emergent phenomena resulting from the complex temporal dynamics of distributed neural systems. To reduce psychology to neuroscience means, to me, to explain/understand/account for psychological behavior in terms of neural behavior. Psychology does not disappear because there is a combinatorial problem: A given psychological phenomena arises out of some combination (in space and time) of neural functioning. If one does not have clear and quantitatively defined "higher level" concepts, one will not know what is an appropriate combination of "lower level" concepts (themselves highly dynamic) that relate best to the behavior under investigation. Even in physics, this is the case. One must have defined concepts, such as temperature, entropy, resistivity, in order to develop a theory that accounts for them and their interrelationships in molecular terms. G&S seem to agree with this (e.g., in their discussion of color opponency), but here is where their terminology undermines a realistic understanding of how this must be done. Psychology is no more a "second-rate" or "place-holder" science relative to neuroscience (sect. 1) than thermodynamics is a place-holder science relative to physics. When dealing with macroscopic materials at temperatures that are neither too high nor too low, a thermodynamic (as opposed to a molecular) description is both appropriate and useful. The same would be true for cognitive functioning and neuroscience. Although cognitive psychology does not disappear, by itself it is inadequate. Indeed, a major challenge comes in developing a framework (such as statistical mechanics in physics) that can bridge these domains. The importance of having a neuroscientific explanation of psychological phenomena is that, although competing psychological theories may equally well account for behavioral phenomena, perhaps only one (or a few) of these theories can connect with the underlying neuroscientific domain, thus allowing a pruning of potential theories to occur.

I also found particularly disappointing G&S's apparent lack of appreciation of how cognitive neuroscience in the last few years has shifted its conceptualization of the way neural systems are related to behavior. Many of their examples and quotes represent an old-fashion neuroscience that tried to understand cognitive behavior based on data obtained from examining one "object" at a time (that object may be a neuron or a brain region). A marked change has come about because of the relatively recent availability of experimental methods that can acquire neural data simultaneously from multiple functional "entities": multiunit electrophysiological recordings in nonhuman species (see, e.g., Wilson & McNaughton 1994) and especially functional neuroimaging (PET, fMRI, EEG, and MEG) in humans (for reviews, see Horwitz 1998 and Simpson et al. 1995). Applying computational modeling to these kinds of data provides a powerful way to relate neuronal activity in multiple interacting brain regions to cognitive behaviors. This, I believe, is the way in which we will be able to understand how neural functioning mediates cognition.

For example, we (Tagamets & Horwitz 1998) recently developed a large-scale, neurobiologically realistic computational model that can perform a delayed match-to-sample (i.e., working memory) visual discrimination task. The task consists of the presentation of a shape, a delay, and the presentation of a second shape; the model must decide if the second stimulus is the same as the first. In this model, we have multiple brain regions, with interregional connectivities (both feedforward and feedback) based on primate neuroanatomical data. The excitatory elements of this model have simulated neuronal activities resembling those found in electrophysiological recordings from monkeys as they perform similar tasks (see, e.g., Funahashi et al. 1993). Furthermore, we can use this model to simulate a functional neuroimaging (PET) study, and the simulated PET activity has values similar to those found in a human PET working memory study (Haxby et al. 1995). Here we have a first step toward relating the dynamic behavior of multiple neuronal populations to human functional neuroimaging to cognitive behavior. All levels of this analysis (neuronal, brain, behavior) play fundamental, but hierarchically arranged, roles. Large-scale, neurobiologically realistic models have been developed by others to address different questions relating neural activity to cognitive behavior (see, e.g., Dominey & Arbib 1992; Tononi et al. 1992).

The construction of such large-scale neural models is hardly "trivial," and their developers would often suggest that such models have succeeded in "explaining" (at least to a first approximation) a specific aspect of psychology in terms of neural activity, consistent with the "radical" neuron doctrine. However, psychological notions were used extensively, and the effect of the work is not to eliminate these notions, or downgrade them, but rather to understand them in terms of the underlying neuroscientific framework. Thus, by dichotomizing this continuous conceptual framework, G&S have misconstrued the way cognitive neuroscience is actually carried out.

A slightly radical neuron doctrine

Frank Jackson

Philosophy Program, Institute of Advanced Studies, The Australian National University, Canberra, ACT 0200, Australia. **frank.jackson@anu.edu.au**
coombs.anu.edu.au/Depts/Rsss/Philosophy/People/Jackson/index.html

Abstract: The element of truth in behaviorism tells us that some versions of a radical neuron doctrine must be false. However, the representational nature of many mental states implies that neuroscience may well bear on some topics traditionally addressed by philosophers of mind. An example is the individuation of belief states.

One can accept that mental states are brain states and that the brain is a "neuronal" machine without accepting that a complete neurobiological or neuroscientific account of us would deliver all there is to say about our psychology. One reason is that many psychological states get their psychological nature in part from how they connect their subjects into their environments. One does not have to be a card-carrying behaviorist to grant that there is a conceptual link between what subjects believe and desire and the way in which these states move subjects through their worlds. However, connections to environments are not delivered by neurobiology or neuroscience alone. Although these connections are mediated by subjects' neuronal natures, one could stare forever and a day at subjects' neuronal goings on without being able to work out what they believe and desire. To work out what they believe and desire, one must know about typical environmental causes and effects in normal circumstances of the neuronal states one is staring at. However, knowledge about environmental connections is not knowledge solely about neurobiology and is not knowledge frameable in terms of the concepts of neuroscience alone. For this

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

reason (and others), I agree with Gold & Stoljar (G&S) in rejecting some radical neuron doctrines. Nevertheless, I think that neuroscience might tell us highly interesting things about our mental natures of a kind that have traditionally engaged the attention of philosophers of mind. Neuroscience might do more than give interesting information about the causal underpinnings of the various mental states, about what realizes the various key functional roles, and the like. It might give information bearing on the individuation of mental states. To this extent, we should accept a slightly radical neuron doctrine.

Mental states such as belief are essentially representational; they serve to represent the world as being a certain way. This is how beliefs get to be true or false: true if things are the way they represent them, false otherwise. In this respect, they are putative bearers of information about our environment (and also about our internal states, as when we believe that we are about to be sick, but I will focus the discussion on outward-looking belief states). Hence, if beliefs are neuroscientific states of heads, then these neuroscientific states must be putative bearers of information about the environments of the heads they are located in; they must be "traces" of these environments; they must be states that somehow code for these environments. Accordingly, what we can very properly expect from future neuroscience is an account of the way brains carry the information, an account of the neuronal coding system. Indeed, many cognitive scientists are already placing their bets on the shape of the account that will emerge. This means that the findings of neuroscience may well bear on the individuation of certain of our mental states.

Consider, for example, my perceptual belief that there is an equilateral triangle in front of me impinging on my sense organs, and my perceptual belief that there is an equiangular triangular in front of me impinging on my sense organs. If it turns out that there is just the *one* neuroscientific state that carries the information equally that there is an equilateral triangle in front of me impinging on my sense organs and that there is an equiangular triangle in front of me impinging on my sense organs, a single state which can be reported by the brain's "language module" either by using the word "equiangular" or by the word "equilateral" in somewhat the way that polar and Cartesian coordinates can report the very same location of a simple point on a graph, then there is just one state with two ways of being reported in the language of folk psychology. When we are counting states, we count one. (However, when we are counting *contents,* we might [*might*] want to insist that there are two contents, that is, that there is one state with two contents, like a single sign that tells us two things at once.) However, if it turns out that our brains are like English in having two codings, one which has, say " . . . lateral" where the other has " . . . angular," or their "brain" equivalents, there are two belief states that may or may not have different contents.

Or perhaps – as many have argued, or at least seriously entertained – it will turn out that our brains have a very holistic way of carrying information about our environments, more like the way in which maps and hologram negatives carry information than the way in which sentences do. In this case, it would be arguable that we should regard talk of individual beliefs as best understood as talk of the individual *sentences* that give one or another aspect of the whole, detailed way that some single belief state or system of belief represents things as being (see, e.g., Lewis 1994, pp. 412–31). Think of the way in which one's current perceptual experience represents how things are in front of one. It is plausible that the various sentences one produces to capture aspects of how that experience represents things as being do not, in any interesting sense, carve at the representational joints of that experience. Perhaps belief is like that quite generally.

To this limited extent, I agree with one theme in the writings of the Churchlands that G&S discuss. Neuroscience might have quite radical implications for the ontology of mind – though, unlike the Churchlands, I would prefer to describe these implications as exciting discoveries about the nature of certain mental states, not the discovery that these mental states do not exist.

The "trivial neuron doctrine" is not trivial

Dale Jamieson

Environmental and Technology Studies Program and Department of Philosophy, Carleton College, Northfield, MN 55057.
djamieson@carleton.edu **www.dir.ucar.edu/esig/HP_dale.html**

Abstract: I argue that the trivial neuron doctrine as characterized by Gold & Stoljar is not trivial; it appears to be inconsistent with property dualism as well as some forms of functionalism and externalism. I suggest that the problem is not so much with the particular way in which Gold & Stoljar draw the distinction as with the unruliness of the distinction itself. Their failure to see this may be why they misunderstand the views of the Churchlands.

According to Gold & Stoljar (G&S), the trivial neuron doctrine (TND) is the view that "a successful theory of the mind will be a solely cognitive neuroscientific theory" (sect. 2.2.1). They go on to identify three components of the TND: materialism, naturalism, and an openness about exactly "which concepts will feature in the successful theory of the mind." Although the TND is trivial in that it " . . . expresses little more than a commitment to an explanation of the mind by science. . . ." G&S point out that the TND is inconsistent with some antiscientific views of the mind such as substance dualism and social constructivism. In fact, however, the TND is also inconsistent with property dualism (PD) and probably with some versions of functionalism and externalism. Because these are widely held doctrines, any theory that is inconsistent with them cannot be correctly characterized as trivial.

PD is (roughly speaking) the view that, whereas only physical substances may exist, they may have nonphysical as well as physical properties. Qualitative properties are the usual candidates for nonphysical properties. If the TND is true, then PD is false. According to G&S, the TND's materialism commits it to the view that "mental phenomena are neural phenomena," but neural phenomena are physical, whereas PD holds that some mental phenomena are not.

Some versions of functionalism (e.g., antipodean functionalism) hold that conceptual analysis of various platitudes about mental states plays an important part in understanding the mind, because it is conceptual analysis that gives content to mental terms whose referents are then identified with physical states. However, the naturalism of the TND commits it to the view that the "understanding of this phenomenon [the mind] will derive from science." Because conceptual analysis of various platitudes about the mind is clearly not a scientific activity, it would seem that any theory that holds that such analysis is central to understanding the mind is inconsistent with the TND.

Finally, although this is less clear, it appears that externalism about mental content may also be inconsistent with the TND. Externalism (roughly speaking) holds that some mental states are in part constituted by facts, properties, or relations that are not internal to the individuals who instantiate mental states. The TND is committed to the view that "mental phenomena are neural phenomena." Insofar as neural phenomena would seem to be internal, it would appear that the TND is inconsistent with externalism.

What has gone wrong? I believe that the reason why the TND turns out to have more far-reaching consequences than G&S suppose is not that there is any particular problem peculiar to their characterization of the TND but rather the inherent difficulty in distinguishing the TND from the strong neuron doctrine (SND). The distinction between these two doctrines is, of course, at the heart of the target article. The linchpin of the distinction is supposed to be the difference between cognitive neuroscience, whose resources the TND is willing to draw upon, and (merely) biological neuroscience, which is all that is available to the SND. However, the very tenability of this distinction is part of what is up for grabs in recent discussions of the mind. It is not as clear as it once appeared to be what cognition consists in or what resources are required to manifest it. The failure to appreciate this is part of the reason why G&S misunderstand the views of the Churchlands.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

The Churchlands (1994) write that:

> In general terms we already know how psychological phenomena arise: They arise from the evolutionary and ontogenetic articulation of matter, more specifically, from the articulation of biological organization. We therefore *expect* to understand the former in terms of the latter. The former is produced by the relevant articulation of the latter. (p. 48)

G&S find a logical fallacy in this passage: The first sentence asserts the TND, and from this the Churchlands fallaciously infer the SND. However, it is tendentious to read the "therefore" in the second sentence as a sign of deductive inference rather than as marking the transition to an expectation about future knowledge on the basis of what we now know. The Churchlands say that what they are attempting in this paper is only " . . . to rebut the counsel of impossibility . . ." regarding " . . . the possibility of reducing psychology to neuroscience . . ." and trying " . . . to locate the reductive aspirations of neuroscience in a proper historical context" (p. 53). In the passage that G&S quote, the Churchlands are citing reasons for supposing that psychology will be eliminated in favor of neuroscience; the Churchlands then go on to discuss what they call "contraindications." What is important about this misreading of the Churchlands is that it brings out an assumption that G&S begin with that is not universally shared. They assume that the distinction between cognitive versus (merely) biological neuroscience is clear and unproblematic, though (revealingly) it begins to slip away when they try to characterize it. However, some would say that such a distinction is an unwieldy basis from which to project important claims. The scope, force, and tenability of this distinction is part of what is at stake in the wide open discussion of the mind that is now underway.

"Mind is brain" is trivial and nonscientific in both neurobiology and cognitive science

J. Scott Jordan

Department of Psychology, Saint Xavier University, Chicago, IL 60655.
jordan@sxu.edu

Abstract: Gold & Stoljar reveal that adherence to the radical neuron doctrine cannot be maintained via appeals to scientific principles. Using arguments from (1) naturalism and materialism, (2) unification, and (3) exemplars, it is shown that the "mind-is-brain" materialism explicit in the trivial version of the neuron doctrine ultimately suffers the same theoretical fate. Cognitive science, if it is to adopt an ontology at all, would be better served by a metaphysically neutral ontology such as double-aspect theory or neutral monism.

Gold & Stoljar (G&S) do an admirable job of revealing the ambiguous nature of the neuron doctrine and then parsing it into its trivial and radical connotations. I am concerned, however, that the success of their arguments against the radical version will be misconstrued as a success of the trivial version. Using the same means employed by G&S (i.e., arguments from naturalism and materialism, arguments from unification, and arguments from exemplars), I will attempt to demonstrate that adherence to mind-is-brain materialism is just as nonscientific in the trivial version of the neuron doctrine as it is in the radical version.

Naturalism and materialism. Although being a scientist reveals a certain commitment to naturalism, a commitment to naturalism neither reveals nor dictates a commitment to materialism. On the contrary, the making of strong ontological commitments is somewhat inconsistent with naturalism. What matters most to the naturalist is that theories of natural phenomena should be based on science and that the statements made in those theories should be based on replicable phenomena. One goes against the naturalist's creed when one clings to the materialist assumption that the ontological basis of the phenomenal world resides in a transphenomenal *material* reality, because such an assumption cannot be empirically tested. By definition, only one of the two (i.e., the phenomenal world) is empirical. The assumed existence of the other (i.e., transphenomenal *material* reality) is neither fact nor testable theory. It is, for all scientific purposes, superfluous conjecture that is derived from replicable phenomena and can be adhered to only via tenacity.

In cognitive science, a commitment to materialism has even graver consequences, for the materialistically grounded notion that mind is brain restricts the theoretical space to notions of material cause and phenomenal effect. From the naturalist's perspective, accepting such conceptual restrictions simply on the basis of superfluous conjecture is unnecessary and counterintuitive. If a cognitive scientist chooses to assume anything regarding transphenomenal reality, it should be the least restrictive ontology available, something along the lines of the double-aspect theory referred to by G&S, or the neutral monism espoused in Bertrand Russell's (1970) event ontology.

Unification. Another consequence of mind-is-brain materialism for cognitive science is an unnecessarily restricted view of potential forms of scientific unification. For example, the mind-is-brain notion leads one to assume that unification, if it is to take place at all, will involve a reduction of psychological concepts to concepts of neurobiological articulations of matter. Adopting a double-aspect or neutrally monistic ontology, however, allows one to entertain the even more radical notion that the concepts "psychological" and "material" may one day find themselves unified within a more parsimonious family of concepts. G&S make this same point in their arguments against the radical neuron doctrine, but they fail to mention that these arguments are just as applicable to the mind-is-brain materialism inherent in the trivial neuron doctrine. Given that materialistic and double-aspect ontologies cannot be discriminated via empirical tests, the naturalist interested in adopting an ontology should adopt the one that serves to retain the maximal number of theoretical possibilities.

Exemplars. An additional consequence of mind-is-brain materialism is that it unnecessarily restricts the types of explanations cognitive science can generate. For example, a mind-is-brain explanation of color phenomena is restricted to the assertion that color exists within the brain. If one backs away from this position by claiming that color phenomenology involves a variety of nonbrain material essences (e.g., electromagnetic radiation and/or material objects in an environment), one has, essentially, retreated from mind-is-brain materialism. The point is that even a materialistic account of color phenomenology cannot appeal simply to the brain. Thus, mind-is-brain materialism is, at best, untenable and, at worst, terribly misleading. When this is coupled with the fact that distinctions between ontologies cannot be made on the basis of empirical tests, and hence can be adhered to only for nonscientific reasons, it should be quite clear to the cognitive scientist that a metaphysically neutral ontology such as double-aspect theory or neutral monism, not materialism, is preferable. Betting naturalists keep their options open.

Radical explanations, but trivial descriptions

Claus Lamm

Brain Research Laboratory, Department of Psychology, University of Vienna, A-1010 Vienna, Austria. **claus.lamm@univie.ac.at**

Abstract: A thorough distinction between explanatory and descriptive concepts reveals a radical explanatory and a trivial descriptive doctrine in current neuroscientific research. The explanatory approach examines the neuronal substrates of the mind, whereas the descriptive one deals only with its correlates.

Most commentators will agree that one of the main merits of Gold & Stoljar's (G&S's) target article is that it points out a lack of broad theoretical and philosophical considerations in the mind-related neurosciences. Particularly in sections 6.1 and 6.2, but also in the

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

provocative statement that "it is only a slight exaggeration to say that we are almost completely ignorant about how the brain produces mental life" (sect. 1.3, para. 2), G&S make it explicit that we are currently on much shakier ground than some recent technical developments, research results, books (e.g., Posner & Raichle 1994), and media coverage might suggest.

G&S argue quite stringently that there is not enough support for a so-called radical neuron doctrine. A more thorough distinction between explanatory and descriptive concepts, however, may reveal the existence of a radical explanatory and a trivial descriptive neuron doctrine in the contemporary mind-related neurosciences. That G&S do not make this distinction becomes evident when one compares their quotations from the proponents of the radical doctrine to their objection to them. Whereas the quotations almost always include the concept of explanation or understanding and also explicitly use these terms (e.g., Churchland & Sejnowski 1992, pp. 3, 239; Crick 1994, p. 7; Snyder 1996, p. 1), G&S's "objection is only to the view that the best *description* . . . will be entirely neurobiological" (sect. 3.2, para. 1, emphasis added).

The aim of the radical *explanatory* approach is to reveal the necessary and sufficient neuronal conditions for the mind, that is, to find the neuronal *substrate* of the mind. Necessary and sufficient mean that such explanations make explicit all steps that are involved in some psychological function (e.g., learning) *on a neuronal level.* Thus, they are radical in G&S's sense. However, this does not mean that terms used in psychology or other behavioral sciences might not be found in the explanation. On the contrary, they must be, because the "thing" to be explained must be referred to. Borrowing from Marr's (1982, p. 27) suggestion that it is inappropriate to understand bird flight by studying only feathers, it is impossible to explain learning by describing only the activities of neurons without referring to the behaviorally overt processes of learning as well. An example of a neuroscientific explanation is Kandel and coworkers' (see, e.g., Kandel & Schwartz 1982) detailed report of the neuronal processes that underlie the phenomenon of a formerly irrelevant stimulus (weak tactile stimulus to the siphon of *Aplysia*) resulting in a gill-withdrawal reaction. Of course, this explanation does not cover the whole spectrum of what psychology calls "classical conditioning," and it is not even necessary to relate the explanation to this theory. It explains only the result of the repeated contiguous presentation of two formerly unassociated stimuli.

In most cases, the precursor to the explanatory approach will be the descriptive one (cf. Reber 1985, p. 191), an approach based on neuroscientific plausibility that at most reveals the sufficient, but not the necessary, neuronal conditions of a psychological function. The descriptive approach analyzes only the neuronal *correlates* of the mind and is trivial in that it is the one that the majority of neuroscientists, and especially cognitive neuroscientists, must currently choose. It is to be chosen when some phenomena that are known on a behavioral level cannot yet be explained or even observed in detail on the neuronal level. One example from descriptive neuroscience is again Kandel and colleagues' work on learning mechanisms in *Aplysia.* They were able to explain in detail the behavioral association of two stimuli by contiguity, but they have not yet been able to explain or observe some of the more complex and perhaps more fundamental aspects of classical conditioning, such as the role of informational content of the unconditioned stimulus (see sect. 5.3.5). Nevertheless, one can easily theorize about its neuronal basis, which might result in an explanation of the role of informational content. As long as this explanation is not found, however, one must rely on description, with psychological and neuroscientific accounts of the phenomenon alternating, neither of them dominant.

Until now, and even with the rapid technical development in the field of behavioral neuroimaging at the close of the "decade of the brain," we are still far from purely neuronal explanations of cognition and behavior. Neuroimaging techniques such as PET and fMRI might yield more detailed descriptions of what is going on in the brain during cognitive processing, providing an enormous amount of exciting new data. However, they give access only to neuronal correlates of the cognitive processes in question (see also Sarter et al. 1996 and multiple book review of Posner & Raichle's *Images of Mind BBS* 18(2) 1995), and other disciplines, such as psychology and computational modeling, are still necessary to explain the neuroimaging data themselves. This might be one reason why Michal Gazzaniga, one of the founders of cognitive neuroscience, is rather cautious in formulating the present aim of his discipline as "figuring out how the mind arises from the brain" (Waldrop 1993, p. 1807) or "how the brain enables the mind" (Gazzaniga 1995, p. xiii) and only sees the future of his field in "a science that truly relates brain and cognition in a mechanistic way." Whether or not we will realize this future some day, I agree with G&S (sect. 5.4.3, para. 5) that the better bet is the descriptive approach.

ACKNOWLEDGMENTS

I am grateful to Judith Glueck and Oliver Vitouch for their comments on an earlier version of this commentary.

A more substantive neuron doctrine

Joe Y. F. Lau

Department of Philosophy, The University of Hong Kong, Hong Kong.
jyflau@hkusua.hku.hk **www.hku.hk/philodep/joelau**

Abstract: First, it is not clear from Gold & Stoljar's definition of biological neuroscience whether it includes computational and representational concepts. If so, then their evaluation of Kandel's theory is problematic. If not, then a more direct refutation of the radical neuron doctrine is available. Second, objections to the psychological sciences might derive not just from the conflation of the radical and the trivial neuron doctrines. There might also be the implicit belief that, for many mental phenomena, adequate theories must invoke neurophysiological concepts and cannot be purely psychological.

In presenting the radical neuron doctrine, Gold & Stoljar (G&S) did not explicitly say whether computational and representational concepts (CRCs, for short) fall within their definition of biological neuroscience; but this is important because these concepts seem to be indispensable in understanding the function of neural mechanisms. Without them, we cannot understand how neurons contribute to information processing in the brain. As a matter of fact, even the Churchlands appeal to notions such as content-addressable memory, distributed representations, parallel processing, and vector transformation in articulating their favorite research program. Such concepts obviously cannot be reduced to neurophysiology, however, as they can also apply to nonbiological systems. Thus, if CRCs are indeed indispensable, and they fall outside biological neuroscience, then this is already sufficient to refute the radical neuron doctrine.

Perhaps G&S meant to include CRCs within biological neuroscience. However, such a move is likely to weaken their argument that Kandel's theory of learning cannot provide a reduction of the concept of classical conditioning. According to G&S, the current conception of classical conditioning involves the learning of relations among represented events. However, this involves the notion of information about relations that they think cannot be captured in Kandel's theory. This might be so, but the issue is whether biological neuroscience in principle has the resources to fill the gap. Insofar as CRCs are ideally suited for capturing informational concepts, proponents of the radical doctrine might reply that Kandel's theory (or an improved version) can provide a reduction of classical conditioning when embedded within a suitable computational framework, and this enriched theory can still be part of biological neuroscience in the broad sense. Whether the radical neuron doctrine is true on this reading would then depend on whether there are psychological concepts that cannot be reduced to CRCs plus other concepts in biological neuroscience. I think that there are indeed many such concepts, but this is not the place to go into the arguments.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

A related issue arising from G&S's discussion concerns the relationship between psychological and neurophysiological theories. G&S seem to think that the latter can at most provide implementations of the former, and they illustrate their point using the theory of color opponency and David Marr's theory of vision. A common feature of both examples is that there is a level of psychological theory that can be specified independently of neural implementation. In the first case it is the theory of the opponent character of color perception; in the second case it is a theory of what the visual system computes and why. Interestingly enough, however, Marr himself cautions that the distinction between computational and implementational theories might not be applicable to all problems of biological information processing. He says that "this can happen when a problem is solved by the simultaneous action of a considerable number of processes, *whose interaction is its own simplest description"* (Marr 1977, p. 38 [his emphasis]). If I understand him correctly, I think his point is that in such situations, which he calls "Type II" situations, it might be impossible to find an informative abstract description of what a system does without mentioning the complex mechanisms involved.

The relevance of Marr's remark is that it raises the following possibility: There might be many mental phenomena for which it is impossible to devise informative and explanatory theories that are purely psychological and that do not make use of neurophysiological concepts. Let the "substantive neuron doctrine" be the claim that this possibility does in fact obtain. Of course, even if this doctrine were true, it would not vindicate the radical neuron doctrine, insofar as the mixed theory can contain irreducible psychological concepts, but this substantive doctrine is not trivial either; it has the methodological consequence that for some mental phenomena it would be misguided to try to develop a purely psychological theory.

The point is not just that one has to keep in mind the issue of neural implementation when devising psychological theories for these phenomena. Rather the claim is that one cannot begin to formulate an adequate theory without explicitly bringing in neural details, "getting one's hands dirty" as it were. It seems to me that a lot of the rhetoric directed against the psychological sciences might have to do with the implicit acceptance of this substantive doctrine and not just the conflation of the radical and the trivial doctrine.

This is one way to interpret what the Churchlands have in mind when they criticize "autonomous psychology" (McCauley 1996, p. 220). They give the example that the structure of the periodic table remains a mystery until quantum mechanics enter into the picture. Likewise, the suggestion might be that many distinctive features of the mind can be explicated only if we bring in neurophysiological findings. Whether this is true is of course an empirical matter. There can be no *a priori* route to the conclusion that, say, theories of syntactic principles must somehow bring in neurophysiological concepts if they are to be viable. As with the rest of science, the ultimate justification for any particular approach lies in its success, but, whatever the case may be, on this interpretation we need not see those who defend the neuron doctrine as defending a view that either has no defense or that needs none.

Supervenience and qualia

Ken Mogi

Sony Computer Science Laboratory, Higashigotanda, Shinagawa-ku, Tokyo, 141-0022 Japan. **kenmogi@csl.sony.co.jp**
www.csl.sony.co.jp/person/kenmogi.html
www.quali-manifesto.com

Abstract: The privileged position of neural activity in biological neuroscience might be justified on the grounds of the nonlinear and all-or-none character of neural firing. To justify the neuron doctrine in cognitive neuroscience and make it both plausible and radical, we must consider the supervenience of elementary mental properties such as qualia on neural activity.

The assumption that neurons are the appropriate level of description for cortical information processing and mental phenomena in general (the neuron doctrine) is usually regarded as valid. It is important, however, to question once in a while the very foundation and scope of this doctrine, as Gold & Stoljar (G&S) have done.

The ultimate reductionist approach to cortical information processing would only point to physics, and the ultimate level of description would be that of elementary particles. From this perspective, as G&S remark, neurobiology would be only a "local stop" (sect. 4.2), so the privileged status of neurons in today's brain science cannot be derived from reductionism itself.

How then is neural firing the appropriate level of description in neuropsychology? From the dynamics point of view, neural activities are special because of the nonlinearity and all-or-none character of action potential generation. No subneural processes are known at present that show the same degree of macroscopic nonlinearity. In addition, in most cases, synaptic interaction is invoked only when a neuron fires. These are the rationales for treating neural firing as the only relevant explicit variable in cortical information processing. All other variables (including those describing the subcellular processes) can be treated as implicit variables, affecting cortical information processing only through their effect on the eventual neural firing. The reductionist would only have to go as far as neural activity; the rest would be details. Neurobiology might be a "local stop," but it suffices. Treating neural firing as an explicit variable does not necessarily entail a grandmother cell-type coding and is, in fact, a generic assumption behind any model of neural coding. It is in this modern sense that the neuron doctrine (Barlow 1972) should be interpreted.

The rather simplified but effective treatment of cortical information processing in terms of neural activities given above does leave some very important issues unanswered, as G&S rightly point out. The main difficulties are in the field of "cognitive neuroscience" as opposed to "biological neuroscience" (sect. 2.1). Here, there is indeed an "ambiguity" in what the neuron doctrine means (sect. 1.4). If it is claimed that the neuron doctrine is relevant only for the biological neuroscience, fine; it is plausible but not radical. If it is claimed that the neuron doctrine supersedes the psychological sciences as well, then it is surely radical, but does not necessarily sound plausible. What is the neuron doctrine really supposed to mean in this view?

In my interpretation, the ambiguity could be resolved by considering the "supervenience" of mental events on neural activities. Davidson (1970) introduced the concept of supervenience thus: "Mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect." To paraphrase, we could hypothesize that there cannot be two events alike in all neural activities but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some neural activities. This hypothesis does sound plausible, and in this sense it is plausible that mental events should supervene on neural activities. In other words, it should in principle be possible to explain mental events in terms of neural activities only, with no extraneous elements needed.

Qualia (Chalmers 1996) come into the picture here. Qualia are the hallmark of our mental activities, at least as far as conscious mental activities are concerned. It seems plausible to assume that a certain quale is invoked in our mind when a certain pattern of neural firing occurs in the brain. There is certainly the difficult question of comparing the qualia that two individuals have. We cannot ever be sure whether the qualia of the red that two subjects have are identical, nor whether such a comparison is meaningful at all. However, it does seem to be plausible that once we have a specific neural firing pattern in individual subjects' brains they will have a certain quale corresponding to that neural activity. In this sense, qualia would supervene on the neural activities.

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

Some authors (including myself) have begun inquiring what the mathematical principles involved in this supervenience would be (Mogi 1999; Santosh, in press).

There is indeed an ambiguity in the neuron doctrine at present. This comes in part from the inability or unwillingness of neuroscientists to come forward with any explicit remarks on how mental events might be derived from neural activities. As long as the neuroscientists persist in this reluctance, the neuron doctrine will be unable to supersede the psychological sciences. My suggestion is that we must address seriously the question of the neural correlates of qualia and must try to understand how mental activities actually supervene on neural activities. If this approach is successful, there will be none of the ambiguity in the neuron doctrine that Gold & Stoljar have pointed out. Elucidating the neural correlates of qualia would make neuroscience plausible and radical at the same time.

Neurobiology: Linguistics' millennium bug?

Stanley Munsat

Department of Philosophy, University of North Carolina, Chapel Hill, NC 27599-3125. **munsat@email.unc.edu**

Abstract: Gold & Stoljar pose a dilemma for linguistics should neurobiology win out as the science of mind. The dilemma can be avoided by reestablishing linguistics as an autonomous discipline, rather than a branch of the science of mind. Independent considerations for doing this are presented.

In raising the specter of the consequences should neurobiology become the science of mind, Gold & Stoljar (G&S) have argued that linguistics has a stake in the outcome. "So linguistics faces a dilemma: Either it must be reformulated in neuroscientific terms," (which is probably not possible) "or else it must be judged a place-holder science" (sect. 1.2). However, there is a way to escape the horns of this dilemma. The proposal is radical in terms of the nature of much current theory in linguistics, but from a broader, more philosophical perspective it is not radical at all. The way to escape the dilemma is to separate linguistics as the study of language from linguistics as a science of the mind.[1] (We may even want to adopt the radical convention of using the term "linguistics" for the former only.) After all, there are many disciplines (musicology, mathematics, jurisprudence, economics, political science) whose subject matters are (loosely speaking) human artifacts. They are not eo ipso branches of the science of mind. Chomsky has asserted the necessity of making linguistics a science of the mind, but we should remember that linguistics had a long and fruitful history, delivering much insight into the structure of natural language, before it was declared a science of the mind. It is hard to believe, for example, that we would have to give up our claims about the character of number and gender agreement in English, or take them to be only place-holder claims, because of the success of neurobiology in explaining the mind. A neurobiological theory of how we *master* these features (the nature of underlying structures and processes, of whatever sort they may turn out to be) would indeed look very different from today's cognitive psychology. Should the millenium drive rules-and-representation psychology out of business, or into the status of a place-holder science, that would be too bad for the psychology of language. Linguistics need not worry about it.

This argument for the separation of linguistics from science of mind may strike some as an argument from prudence and, as such, empty of intellectual force. After all, why should linguistics change its way of doing business now on the basis of a purely speculative threat, to wit, radical neurobiology? A concrete example may help show why this return to treating linguistics as an autonomous discipline is a sound strategy.

Agrammatic aphasia is a disorder associated with damage to Broca's area. The collection of symptoms is rather striking; as the name suggests, syntactic functioning is affected, whereas semantic functioning remains relatively untouched. Thus patients will show problems with number and gender agreement, tracking direct versus indirect objects, and handling "contentless" morphemes such as -en and -ing. They have no problems when the morphemes have content such as dis-, un-, or -able. Prepositions are also selectively affected. Patients show no problems with prepositions when they have content, for example, "under the table" as opposed to "on the table" or "above the table," but performance breaks down when the prepositions have no semantic content, as in "under arrest" or "on fire" ("on" here does not contrast with "over" or "underneath"). Note that the *distinction* between the two types of word (often referred to as "content words" vs. "function words") is made on the basis of facts of the *language*, such as whether or not the preposition, as used, belongs to a contrast set. Suppose the question came up of whether a similar distinction could be made for the prepositions in fused verb-preposition structures such as "throw up," "mess up," and "screw up." Is the preposition "up" in these a content word or a function word (of course the answer may not be the same for all of them)? Linguists would rightly complain if one proposed to answer this question by seeing if performance on these prepositions was down for patients with agrammatic aphasia resulting from damage to Broca's area. Linguistic questions must be settled by citing linguistic facts. The question of how our brain or mind works to enable us to operate with such words is another matter.

But if this is true when the mechanisms in question are neurological structures, it is just as true when the mechanisms[2] are rules and representations. The fact that a set of rules adequately describes what is grammatical (and what is not) does not show that those rules are somehow involved in the actual production of sentences by the individual. The rules (supposing there are such) operating in the production of utterances might be completely different from the rules linguists develop to describe the language. For example, there is no reason why there could not be two different rules sharing what is linguistically a single phenomenon, nor is it to the point to claim that the rules by which the utterances are actually produced are "notational variants" of the linguist's rules. If in fact some rules or others are actually powering the production of our utterances, then there is one and only one correct formulation of what those rules are. This is just how it is with devices that operate by following rules. (If a program will not run on your machine, it is of little consolation that it is a "notational variant" of one that will.)

Finally, I offer a parting shot in favor of the neurological approach. Broca's aphasia unites two otherwise disparate linguistic phenomena – syntactic structure and function words. The explanation of why they "break down" together is a mystery for the rules approach but is immediately explained once we see that they have a common neurological underpinning, Broca's area. It might be suggested that rules dealing with syntax and function words, on the one hand, and rules dealing with semantic content, on the other, are "implemented" in different parts of the brain, and so the neurobiological approach gets no explanatory edge in this case. But why should that be? That is, why should rules dealing with different aspects of language (e.g., syntax vs. semantics) be implemented in structurally different parts of the brain? The whole point of thinking of the brain as organized along the lines of a computer is that the explanatory work is done by the sequencing of the code and the basic algorithms that underlie it. However different syntax is from semantics, the tools for encoding the *rules* of syntax and semantics are the same. A line of code is a line of code. But on the neurobiological view (where, as it were, everything is implementation), one would expect that different brain structures might be better suited for handling different processing tasks.

NOTES

1. Indeed psychologists currently distinguish between linguistics and psychology of language.

2. Here, and throughout this commentary, I use the term "mecha-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

nisms" in a broad, theoretically neutral sense. Hence the "mechanisms" of language could turn out to be rules and representations or neurological structures.

Begging the question of causation in a critique of the neuron doctrine

J. Tim O'Meara

Obermann Center for Advanced Studies, University of Iowa, Iowa City, IA 52242. **omeara@anthropology.unimelb.edu.au**

Abstract: Gold & Stoljar's argument rejecting the "explanatory sufficiency" of the radical neuron doctrine depends on distinguishing it from the trivial neuron doctrine. This distinction depends on the thesis of "supervenience," which depends on Hume's regularity theory of causation. In contrast, the radical neuron doctrine depends on a physical theory of causation, which denies the supervenience thesis. Insofar as the target article argues by drawing implications from the premise of Humean causation, whereas the radical doctrine depends on the competing premise of physical causation, the resulting critique of the neuron doctrine amounts largely to begging the question of causation.

Gold & Stoljar (G&S's) argument rejecting the "explanatory sufficiency" of the radical neuron doctrine depends on distinguishing it from the trivial neuron doctrine. This distinction depends in turn on the thesis of "supervenience," or "Humean supervenience," which underpins claims that distinctively mental properties, entities, and events "supervene" on physical brain properties, entities, and events, and furthermore that such supervenient properties, entities, and events have direct explanatory import beyond that of physical brain mechanisms. Humean supervenience thus offers Cartesianists a fallback position that Davidson (1980) labels "anomalous monism," which endorses ontological monism together with causal (and therefore explanatory) dualism. The trivial neuron doctrine is just the neuron doctrine interpreted through anomalous monism.

Humean supervenience depends in turn on Hume's regularity theory of causation, which takes causation to be a matter of relations among events. Under this theory, causal efficacy can be claimed for any property, entity, or event where (1) statements about their occurrence are accepted as true by some reference group and (2) their occurrence is suitably correlated with another type of event (see, e.g., Nagel 1961, pp. 541–44). Once causation is taken to be a matter of relations among events, causal efficacy is automatically accorded to whatever properties or entities are taken to constitute or demarcate those events (see O'Meara 1997).

Humean causation thus implies that relational properties are causal properties in themselves. Such properties include group and qualitative properties as well as function, fitness, order, meaning, disposition, propensity, and such obtuse properties as is "transported to a distance of less than three miles from the Eiffel tower" (Fodor 1994, pp. 690–91). The resulting claims of causal efficacy and direct explanatory relevance for such properties underpins G&S's conclusion that "special sciences" such as psychology and linguistics, which deal with those supervenient or relational properties, are distinct from the physical sciences.

Humean causation carries the further implication that causal laws state regularities of occurrence among types of events; that causal explanations are logical arguments showing that an event happens because its occurrence accords with the lawful pattern governing such events; that causal concepts, laws, and theories occur at different conceptual "levels"; and that explanatory theories might therefore be "reduced" from one level to another by showing that one set of concepts and laws can be derived from the other (see O'Meara, in press). Humean causation is thus a major premise underpinning the target article's claims concerning what counts as explanatory concepts, properties, and entities; what counts as explanatory laws and theories; what counts as legitimate or sufficient explanations; and what counts as successful reduction.

In contrast, quotes from proponents of the radical neuron doctrine show that they are working with a physical theory of causation. According to this theory, it is empirically factual that (1) causation is exclusively a matter of the physical properties of and interactions among what can be characterized loosely as physical entities (recognizing that they have a temporal dimension), (2) causal properties are exclusively physical properties, and (3) causal entities are exclusively physical entities. Distinctively psychological or cognitive properties and entities might still be said to supervene on physical brain properties and entities – and it might be useful sometimes to do so – but these supervenient properties and entities have no direct explanatory import because their referents lack causal efficacy (see O'Meara, submitted). Thus, the radical neuron doctrine is just the neuron doctrine interpreted through a physical account of causation.

Progress is now being made in formalizing a physical account of causation (see Salmon 1984; 1994; Dowe 1992; 1995). Salmon (1984) notes that this physical account commits us to a "mechanistic world view" that is unpopular, but this unpopularity is unjustified because it is usually conceived in terms that are scientifically outmoded (p. 241). Maxwell's electromagnetic theory and Einstein's special theory of relativity show that electromagnetic fields have fundamental physical reality – an important point for neuroscientists – so that a mechanical philosophy remains viable even though a crude materialism, such as that engaged by G&S, remains untenable (Salmon 1984, p. 241).

Contrary to the claim that researchers "confuse" and "conflate" the trivial neuron doctrine with the radical neuron doctrine, that distinction arises only under the premise of Humean causation, which trivializes the neuron doctrine on the one hand and undermines it on the other. Insofar as the target article argues largely by drawing implications about the neuron doctrine from its premise of Humean causation, whereas the radical doctrine depends on a competing premise of physical causation, the authors' critique of the neuron doctrine amounts largely to begging the question of causation. Researchers who accept a physical theory of causation have every reason to accept the radical neuron doctrine as well.

ACKNOWLEDGMENT

I gratefully acknowledge the support services provided by the Obermann Center for Advanced Studies, University of Iowa, during preparation of this commentary.

The neuron doctrine in psychiatry

Christian Perring

Philosophy and Religious Studies, Dowling College, Oakdale, NY 11769.
perring@dowling.edu **www.angelfire.com/ny/metapsychology**

Abstract: Gold & Stoljar's target article is important because it shows the limitations of neurobiological theories of the mind more powerfully than previous philosophical criticisms, especially those that focus on the subjective nature of experience and those that use considerations from philosophy of language to argue for the holism of the mental. They use less controversial assumptions and clearer arguments, the conclusions of which are applicable to the whole of neuroscience. Their conclusions can be applied to psychiatry to argue that, contrary to many researchers' assumptions, the approaches to both understanding and treating mental disorders must be interdisciplinary.

Gold & Stoljar's (G&S's) argument is wonderfully clear. It is noteworthy that their criticism of the radical neuron doctrine does not rely on considerations about the subjectivity of experience or the holism of the mental. The philosophical debates about subjectivity, what it is like to be a bat or a human, and the existence of qualia are controversial and have narrowed the focus on the interrelation between neuroscience and philosophy to some very specialized is-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

sues. The debates about the holism of the mental are also controversial in philosophical circles and have tended to be very abstract, making it difficult to draw specific conclusions from them for neuroscientific research. G&S's argument relies partly on the simple virtue of conceptual clarity about the relation between metaphysics and scientific explanation and partly on insights about scientific methodology. Their argument is powerful for two main reasons: First, it does not rely on philosophically controversial assumptions; second, it has both conceptual and methodological implications with a far wider reach than those of standard philosophical criticisms of neurobiological accounts of the mind.

One can emphatically endorse G&S's call for more philosophy of neuroscience. Philosophy of neuroscience overlaps with another area of increasing prominence, the philosophy of psychiatry (see the summary of the developments in this new area of research in Perring 1998). This commentary discusses the implications of G&S's target article for psychiatry.

In psychiatry, achieving effective treatments of mental disorder has priority over the understanding of the mind. As in the rest of medicine, it is not unusual to have an effective treatment with little understanding of what causes the disorder or how the treatment works. One of the main methods in finding treatments for mental disorders has the following simple format: (1) Find the causes of a mental disorder. (2) Discover how to prevent the initial causes of the mental disorder, or alternatively discover how to intervene in the chain of causes and effects that lead from the first cause to the final effect.

The neuron doctrine states, roughly, that a successful theory of the mind will be a solely neuroscientific one. It seems to follow from this that the causes of mental disorders should be understood solely within a neuroscientific framework. We have to be very careful how we understand this. For example, suppose for the sake of argument that abandonment of a small child by its primary caregivers will predispose that child to clinical depression later in life. The abandonment itself cannot be understood in neuroscientific terms. It might be possible, though, to describe the effects of the abandonment in neuroscientific terms, and presumably this is what the neuron doctrine would hold. The neuron doctrine would never want to deny that events occurring outside of the central nervous system have significant effects on our mental life. It would hold that, when the mind interacts with the world, the mental aspects of that interaction can be understood with neuroscientific theories.

What are the implications of the neuron doctrine for the treatment of mental disorders? Although obvious, it is worth emphasizing that it does not follow that psychiatry should not concern itself with social issues (Perring 1996). As was just noted, it is quite compatible with the neuron doctrine that mental disorders can have causes (such as the abandonment of small children) not describable in neuroscientific terms.

Let us now focus on psychiatric treatments that deal with the individual rather than the social. Take first the trivial neuron doctrine. It holds that a successful theory of the mind, and thus of mental disorders, will use an interdisciplinary approach integrating the physical and biological sciences with the psychological sciences. Thus the approach to discovering treatments for mental disorders will also be interdisciplinary.

To move to the radical neuron doctrine, it holds that a successful theory of the mind, and thus of mental disorders, will be a solely biological neuroscientific theory. Although many researchers in psychiatry would disagree with the radical neuron doctrine, this assumption underlies much research on the nature of mental disorders carried out today. Furthermore, it is increasingly the standard assumption in the popular understanding of psychiatry in the western world, especially the United States: Mental disorders are often described as "brain disorders," with the clear implication that they should be understood in such terms as malfunctions of neurotransmitter receptors and imbalances of neurotransmitter levels in the bloodstream. Furthermore, the "brain disorder" model is often used as an explanation of why the person with the disorder has no control over the behavior associated with that disorder.

Suppose that the radical neuron doctrine were true. It would *not* follow logically that the treatment for mental disorders should necessarily be that of organic psychiatry, for example, pills, neurosurgery, electroshock treatment, or other direct interventions in the workings of the brain. Other treatments, such as talk therapy, might still be as effective as or more effective than organic treatments. It would still be an empirical question which is the most effective form of treatment. Nevertheless, a case can be made that if the real understanding of the mind is to be found in neurobiology and we have some reasonable chance of achieving such an understanding, then ceteris paribus (supposing we can get a handle on what the ceteris might be) we have the best chance of finding effective treatments by investigating organic treatments as opposed to other forms of treatment.

Now we can see G&S's argument that the radical neuron doctrine is controversial and implausible has significant implications both for psychiatric research and also for the popular understanding of psychiatry. It is likely that an underlying assumption of much research into the nature of mental disorders is false. The arguments set forth in sections 3, 4, and 5 point to the need for an interdisciplinary approach to psychiatric research. It is a further corollary that the issue of the responsibility of those with mental disorders for their behavior has not been settled and still requires careful analysis.

Neuroscience and the explanation of psychological phenomena

Antti Revonsuo

Center for Cognitive Neuroscience, Department of Philosophy, University of Turku, FIN-20014 Turku, Finland. **antti.revonsuo@utu.fi**
www.utu.fi/research/ccn/consciousness.html

Abstract: Explanatory problems in the philosophy of neuroscience are not well captured by the division between the radical and the trivial neuron doctrines. The actual problem is, instead, whether mechanistic biological explanations across different levels of description can be extended to account for psychological phenomena. According to cognitive neuroscience, *some* neural levels of description at least are essential for the explanation of psychological phenomena, whereas, in traditional cognitive science, psychological explanations are completely independent of the neural levels of description. The challenge for cognitive neuroscience is to discover the levels of description appropriate for the neural explanation of psychological phenomena.

What are the proper levels of description and explanation in neuroscience, especially when it comes to the explanation of mental phenomena? Before we try to answer this question, it may be illuminating to consider how the explanation of complex biological phenomena is realized elsewhere in the life sciences. The overall idea is that in a biological system there are several different levels of organization. Biological science tries to determine what these levels are and to construct the corresponding levels of description. The mechanistic explanation of a phenomenon at a certain level of description is accomplished by showing how the structure and behavior of the system at a higher level of description can be understood in terms of the parts of the system, their functions and interactions, all of which reside at a lower level of description.

Mechanistic explanation by no means entails what philosophers call "elimination." Although we do have a mechanistic understanding of, say, a living cell, no biologist would deny that cells nevertheless really do exist, nor that they are elementary entities at their own level of organization. Furthermore, mechanistic explanation in biology has little to do with what philosophers call "theory reduction." Biologists do not engage in figuring out what the logical relationships of linguistic representations of biological knowledge might be. In fact, scientific knowledge in biology does

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

not consist of laws that are expressed in linguistic representations. Thus, theory reduction is largely irrelevant to the actual empirical progress made in biology. Instead of laws, biological knowledge is typically expressed in *models* of biological systems. Such models are usually expressed in visualizable diagrams and figures, and their development is often heavily dependent on the available research methods and techniques that allow the visualization of otherwise unobservable biological phenomena (Bechtel & Richardson 1993; Sargent 1996).

In this naturalistic framework of biological explanation, the radical neuron doctrine (RND) can be reinterpreted as the claim that there should be a privileged level of explanation in neuroscience at the level of the basic structural and functional properties of neurons, ensembles, or structures. All higher level (especially psychological) descriptions can be discarded (at least in principle) in favor of descriptions at the privileged neurobiological level. Gold & Stoljar (G&S) are quite correct in pointing out that the RND is highly dubious. One reason for this (not mentioned by G&S) is that explanation elsewhere in biology obviously requires multiple levels of description; different kinds of biological entities and interactions are realized at distinct levels of organization. There is no reason to believe that any single explanatory level should prevail at the expense of all the other ones or that mechanistic explanation should lead to the abandonment of higher level descriptions and phenomena. A similar view is arising in the philosophy of chemistry: Molecular structure is seen as a level of reality ontologically distinct from the lower, microphysical levels (Del Re 1998; see also Guterman 1998).

It would appear rather foolish for neuroscientists to support the RND when explanation elsewhere in the life sciences actually requires multiple levels of description. I therefore doubt that many neuroscientists actually support the RND. Furthermore, I doubt that neuroscientists have in mind anything so loose as what G&S call the trivial neuron doctrine (TND) when they use the term "cognitive neuroscience." G&S (sect. 2.2.1, para. 1) claim that "cognitive neuroscience includes any concept from the psychological or biological sciences" and that "most cognitive scientists believe they are theorizing about the brain" and consequently are "neuroscientists in the cognitive sense" (sect. 3.1, para. 3). It appears, instead, that cognitive neuroscience attempts to extend the strategies of mechanistic biological explanation all the way up to psychological phenomena. This goal is obvious if we look at how the pioneers of cognitive neuroscience characterize their field:

> Cognitive neuroscience is an attempt to understand how cognition arises from brain processes; the focus is on the brain, as the term "neuroscience" implies. We don't want to separate the theory of information processing from the theory of the brain as a physical mechanism. . . . A complete cognitive neuroscience theory would specify . . . how each process is instantiated in the brain, and how brain circuits produce the input/output mappings accomplished by each process. . . . Given these goals, it seems clear that cognitive neuroscience must move closer to neurobiology. But it will simply not become neurobiology. (Kosslyn, in Gazzaniga 1997, pp. 159–60)

> At some point in the future, cognitive neuroscience will be able to describe the algorithms that drive structural neural elements into the physiological activity that results in perception, cognition, and perhaps even consciousness. . . . The future of the field, however, is working toward a science that truly relates brain and cognition in a mechanistic way. . . . The science built up to understand how the brain enables the mind has come to be called cognitive neuroscience. (Gazzaniga 1995, p. xiii)

Cognitive neuroscience sees psychological levels ("cognition") as the higher levels of description, to be explained by referring to the neural and neurocomputational mechanisms residing at the lower levels. Psychological phenomena are not explanatorily autonomous, but neither are they eliminable, just as cytology is neither eliminable nor autonomous in relation to biochemistry and molecular biology. The cognitive neuroscientific view of explanation is radically different from the standard computationalist and representationalist views in cognitive science, which see psychology as residing at an explanatorily autonomous level:

> A better understanding of the mind is not to be obtained by knowledge – no matter how detailed or precise – of the biological machinery by means of which the mind does its job. (Dretske 1995, p. xiv)

> The view that cognition can be understood as computation is ubiquitous. . . . In studying computation it is possible, and in certain respects essential, to factor apart the nature of the symbolic process from properties of the physical device in which it is realized. (Pylyshyn 1990, pp. 18, 29)

The central disagreement between cognitive science and cognitive neuroscience in the explanation of psychological phenomena is whether or not the explanations can be construed at a level of description where no reference to neural phenomena is necessary. Traditional cognitive scientists claim that no reference to the neural level is really required. Therefore, it seems quite peculiar to call cognitive scientists "neuroscientists in the cognitive sense" (sect. 3.1, para. 3). According to cognitive neuroscientists, *some* neural levels of description at least are essential for the explanation of psychological phenomena: Psychology is not entirely autonomous in relation to neuroscience. Cognitive neuroscience entails no commitments either to the RND (which imposes too strict constraints on acceptable levels of explanation) or to the TND (which imposes practically no constraints at all and is consistent with the complete autonomy of psychological explanation and theory).

We have, however, no reason to believe that current neuroscience describes all the levels of organization in the brain that will be explanatorily relevant. Quite the contrary: We have good reasons to believe that there are higher levels of neurophysiological organization that have not yet been empirically uncovered. Moreover, the ontology of cognition remains unclear, too: Neither symbol processing nor connectionism seems to be an entirely satisfactory account of what cognitive phenomena are at bottom (Bechtel 1994), and the notions of "symbol," "computation," and "algorithm" do not name any natural biological phenomena that would exist independently of observers (Searle 1992). Thus, the appropriate levels of organization remain to be discovered both in the mind and on the brain side of the story.

In conclusion, explanation in neuroscience should be seen in the context of biological explanation. This approach reveals that the RND is indeed an untenable position. What most so-called reductionists in neuroscience seem to have in mind is not the RND but, instead, the notion of mechanistic biological explanation across different levels of description. This type of explanation works extremely well within life sciences and should therefore be extended to psychological phenomena through the research program of cognitive neuroscience. The view opposite to this sort of "reductionism" is not the TND but, rather, the widespread belief within behavioral sciences that psychological explanation is essentially autonomous with respect to neuroscience. Functionalism and computationalism, popular varieties of this philosophy, are deeply embedded in traditional cognitive science. Cognitive neuroscientists are thus caught in a theoretical dilemma: On the one hand, they absorb the psychological vocabulary from cognitive science, which emphasizes independence from neural and biological levels of explanation. On the other hand, cognitive neuroscientists want to have strictly mechanistic biological explanations of psychological phenomena. This curious combination of ideas from biological and cognitive sciences leads them to talk about computation in the brain and "algorithms . . . translating structural physiological data into psychological function" (Gazzaniga 1997, p. 160), although it is empirically impossible to discover algorithms or computations in the brain, insofar as these notions do not denote any natural phenomena (Revonsuo 1994; Searle 1992). These explanatory problems reduce to the fact that we simply have not yet found the appropriate levels of description to be used in a

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

full mechanistic biological explanation of psychological phenomena.

ACKNOWLEDGMENT
The writing of this commentary was supported by the Academy of Finland (project 36106).

Neural circuits and block diagrams

J. J. C. Smart
Philosophy Program, Research School of Social Sciences, Australian National University, Canberra ACT 0200, Australia.
jjcs@coombs.anu.edu.au
coombs.anu.edu.au/dept/RSSS/philosophy/people/smart.html

Abstract: This commentary is intended to illuminate Gold's & Stoljar's main contentions by exploiting a favorite comparison, namely, that between biology and electronics. Roughly, and leaving out Darwinian theory and the like, biology is physics and chemistry plus natural history just as electronics is physics plus wiring diagrams. Natural history (even that discovered by sophisticated apparatus such as electron microscopes) contains generalizations, not laws. Psychology and cognitive science typically give more abstract explanations, as do "block diagrams" in electronics, and are less dispensable.

I am largely in agreement with Gold & Stoljar (G&S) and here wish to illuminate and possibly strengthen what I believe to be their main contentions by using a favorite comparison of mine (Smart 1963; 1989), that the relation of physics and chemistry to biology (and I take psychology to be a branch of biology) is comparable to that of physics to electronics. (I am thinking of the central biochemical core of biology, not of Darwinian theory or of statistically and geographically oriented ecology.) Putting it crudely, biology is physics and chemistry plus natural history, and electronics is physics plus wiring diagrams. I shall abstract away from the practical motives of electronics and think of it as an object of intellectual curiosity, as some of us nontechnologists do. Natural history does not give us laws but gives us mere generalizations, exceptions to which need not cause intellectual concern. The generalizations may be more sophisticated than those made by field naturalists: They may be about entities that can be observed only with highly technical apparatus. With a qualification (having to do with "block diagrams"), the generalizations are explained by means of the laws of physics and chemistry. Thus, it is wrong to think of biology becoming a subtheory of physics itself.

Now, let us look at the analogy with electronics. Wiring diagrams are generalizations. Apparatus can go wrong; components can fail. Consider a superheterodyne radio receiver. The distinctive thing about it is a circuit (based on a valve or transistors) that changes a radio frequency signal to an intermediate frequency signal, which can be amplified with greater stability than can a radio frequency one and with greater sensitivity than can an audio frequency one. This circuit is called "a frequency changer" and can be represented as a square in a "block diagram," which will also contain other interconnected but distinct squares, such as for "radio frequency amplifier," "detector," and "audio frequency amplifier" as well as for "intermediate frequency amplifier." Note that it is possible for these entities to overlap in the wiring diagram. A block diagram for a superheterodyne certainly can give some understanding of the working of the receiver. It would not be as detailed an explanation as we would get from a complete wiring diagram. Nevertheless, it would have some advantages, enabling us to see the wood for the trees and also being more general, applying to superheterodynes whose circuits differ, for example, regarding whether there are thermionic valves or transistors. Pursuing the analogy, cognitive science and psychology operate very largely at a "block diagram" level. In the electronics case, the explanatory value persists even in the absence of teleology and in the possible intimate knowledge of the circuitry. Even with this last we are not quite down to physics; "transistor," "transformer," "resistor," and so on are hardly terms of physics proper, nor is "neuron." Engineers regularly treat components as "black boxes," and neuroscientists regularly do the same with neurons. Still, this is not germane to my main point.

Let us therefore think of psychology and cognitive science as largely concerned with block diagrams. The unification of these sciences with neuroscience is thus the bringing together of one level of abstraction with another. Neurology itself contains generalizations of natural history, not laws, as in the originally separate sciences of electricity and magnetism. The unification of these in Maxwell's equations and the invariance of these in special relativity is a poor model for the sort of unification of generalizations and physical principles that we get in biology and in technology. Perhaps what is wrong with the position of the proponents of the neuron doctrine is that unification has to be of the sort we get in physics or else is nothing. I hope that my comparison of the explanatory structure of biology to that of technology (in abstraction from teleology) will shed light on the confusion that Gold & Stoljar diagnose: the conflation of a weak and a strong interpretation of the neuron doctrine. Indeed, neuroscience may always have to rely on block diagrams, because the circuitry of the human brain is thousands of orders of magnitude more complex than that of a radio, and the values of interconnections between neurons are changing all the time. *Metaphysically,* of course, I have no quarrel with the strong neuron theory.

Autonomous psychology and the moderate neuron doctrine

Tony Stone[a] and Martin Davies[b]
[a]Division of Psychology, South Bank University, London SE1 0AA, United Kingdom, and [b]Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom. **stonea@sbu.ac.uk**
martin.davies@philosophy.ox.ac.uk sbu.ac.uk/psycho

Abstract: Two notions of autonomy are distinguished. The respective denials that psychology is autonomous from neurobiology are neuron doctrines, moderate and radical. According to the moderate neuron doctrine, interdisciplinary interaction need not aim at reduction. It is proposed that it is more plausible that there is slippage from the moderate to the radical neuron doctrine than that there is confusion between the radical neuron doctrine and the trivial version.

What does it mean to say that the discipline of psychology is autonomous? Jerry Fodor (1998, p. 9) says that "a law or theory that figures in bona fide empirical explanations but that is not reducible to a law or theory of physics is ipso facto autonomous." F-autonomy is irreducibility. The Churchlands mean more than this by autonomy; for them, to regard a discipline as autonomous is to "try to conduct the affairs of [that discipline] independently of the affairs of its immediate neighbours, both upward and downward in level" (Churchland & Churchland 1996a, p. 220). We take it that this is not just a matter of pragmatic choices about conducting research with limited resources; rather, a discipline is autonomous when it is not governed or constrained by other disciplines. In this sense, C-autonomy is independence. On the face of it, the two doctrines are distinct, and C-autonomy entails F-autonomy.

Consider now the denial of autonomy in each case. Officially, the denial of F-autonomy about psychology is the claim that if there are laws or theories of a genuine empirical science of psychology, they are reducible to laws or theories of physics. We shall suppose that, en route to physics, the laws or theories of psychology would be reduced to laws or theories of neurobiology. This is the aspect of the denial of F-autonomy that we shall focus on: the denial of the autonomy of psychology from neurobiology. The de-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

nial of C-autonomy about psychology is the claim that psychological theories are constrained from above and below: "Theories at different levels quite properly function as ongoing checks, balances, and inspirations for theories at adjacent levels, both up and down" (Churchland & Churchland 1996a, p. 221). In particular, the denial of the C-autonomy of psychology from neurobiology holds that discoveries in neurobiology may put empirical pressure on psychological theories.

These two denials are, we may say, neuron doctrines, though they do not equate precisely with the radical and trivial neuron doctrines that Gold & Stoljar (G&S) distinguish. The possibility of reduction is certainly part of the radical neuron doctrine, but it is also part of the doctrine that "the psychological sciences must be relegated to a second-rate, or place-holder, status" (sect. 1, para. 4) and that psychological theories will be discarded in favor of neurobiological ones. As the Churchlands argue persuasively (Churchland & Churchland 1990), this claim about actual scientific practice does not follow from the claim about reducibility. Thus, the denial of the F-autonomy of psychology from neurobiology is a substantial part of, but not the whole of, the radical neuron doctrine.

The denial of C-autonomy is clearly not radical, but it is not trivial either. The trivial neuron doctrine states only that a successful theory of mind will draw on some or all of the components of cognitive neuroscience; it might draw only on neurobiology or only on psychology or on some combination. Thus, the trivial neuron doctrine would be supported by someone who maintained that neurobiology would play no part at all in a successful theory of the mind. As neuron doctrines go, this is trivial indeed.

We propose to distinguish between a weak and a moderate neuron doctrine. The weak neuron doctrine goes beyond the trivial by saying that neurobiology will be one of the disciplines that together furnish a successful theory of mind. However, this is still consistent with C-autonomy; the weak neuron doctrine allows a picture of cognitive neuroscience as a project in which different aspects of behavior receive explanations from different and independent component disciplines. The moderate neuron doctrine takes the further step of saying that cognitive neuroscience is an interdisciplinary, and not just a multidisciplinary, project. It denies C-autonomy and allows that results in neurobiology could count against a putative cognitive psychological explanation of some aspect of behavior.

Just as the two notions of autonomy are, on the face of it, distinct, so also this moderate neuron doctrine is different from the radical neuron doctrine. The moderate neuron doctrine seems to capture something of the idea that cognitive psychological and neurobiological theories coevolve (P. S. Churchland 1986), but Churchland links the idea of coevolution with something else (p. 284): "The discoveries and problems of each theory may suggest modifications, developments, and experiments for the other, and thus the two evolve towards a reductive consummation." As against this, we do not regard "reductive consummation" as the inevitable result, or even as the desired endpoint, of coevolution or interdisciplinary interaction. Interaction without reduction seems to be an allowable, and even attractive, option.

Among those theorists who would accept the moderate neuron doctrine but without any commitment to the reducibility of cognitive psychological theories, laws, or categories are those who claim that, within cognitive neuroscience, the functional (i.e., cognitive psychological) level has priority over the neurobiological level. Part of the priority idea is this (Coltheart & Langdon 1998, p. 150): "It can be very hard to understand what a system is actually doing if one's only information about it is a description at the physical-instantiation level" (cf. P. S. Churchland 1986, p. 373: "Neuroscience needs psychology because it needs to know what the system does"). Another part is that neurobiological theories may be "conceptually dependent" on cognitive psychological theories (Coltheart & Langdon 1998, p. 149), and, in practical terms, the development of a cognitive psychological theory may abstract from debates within neurobiology even while it is acknowledged that the theory would have to be rejected were there no neurobiological story consistent with it. (See Young 1998, p. 44, for this point applied to a dual-route theory of face processing.)

Perhaps defensible claims about the theoretical and practical priority of the cognitive level sometimes tip over into something more extreme and implausible, namely, the claim that neurobiology is strictly irrelevant to cognitive psychological theorizing. However, even the strong priority claims of Mehler et al. (1984) are consistent with the in principle answerability of cognitive psychological theory to neurobiological data (see also Shallice 1988, p. 214). There is certainly nothing inevitable about a shift from the priority idea to an assertion of strict disciplinary independence. On the contrary, the priority idea seems to be consistent even with the reducibility of cognitive psychology to neurobiology.

Like G&S, we are not convinced of the truth of the radical neuron doctrine in either its theoretical ("reduce") or practical ("discard") aspect. We do, however, think that the moderate neuron doctrine is plausible. Cognitive psychology is constrained by neurobiology because neurobiology tells us about the mechanisms in virtue of which cognitive psychological generalizations are true (Fodor 1989). In practice this is constraint without government; challenges and insights flow in both directions.

We can hypothesise that some arguments for the radical neuron doctrine involve a degree of slippage from interaction to reduction, but G&S's diagnosis of where the arguments are apt to go wrong is different. Their hypothesis, which figures especially in their discussion of the argument from naturalism and materialism, is that advocates of the radical neuron doctrine confuse it with the trivial version. Given that the trivial neuron doctrine allows that neurobiology has no part to play in a successful theory of the mind, whereas the radical doctrine asserts that only neurobiology has any part to play, this is not an easy confusion to make.

The Churchlands' neuron doctrine: Both cognitive and reductionist

John Sutton

Department of Philosophy, Macquarie University, Sydney, NSW 2109, Australia. **jsutton@laurel.ocs.mq.edu.au**
www.phil.mq.edu.au/staff.htm

Abstract: According to Gold & Stoljar, one cannot consistently be both reductionist about psychoneural relations and invoke concepts developed in the psychological sciences. I deny the utility of their distinction between biological and cognitive neuroscience, suggesting that they construe biological neuroscience too rigidly and cognitive neuroscience too liberally. Then, I reject their characterization of reductionism. Reductions need not go down past neurobiology straight to physics, and cases of partial, local reduction are not neatly distinguishable from cases of mere implementation. Modifying the argument from unification as reduction, I defend a position weaker than the radical but stronger than the trivial neuron doctrine.

Gold & Stoljar (G&S) allow "biological neuroscience" to *include* study of the function as well as the structure of neuronal ensembles (sect. 2.1). But they think that invocations of function in actual neurobiological explanations already invoke nonbiological concepts, so that explanations of causal mechanisms fulfilling those functions are *not* purely neurobiological. Because even an apparently physiological notion such as the reflex is "highly theoretical," G&S deny that it is a legitimate construct of physiology alone (n. 40). It is as if the fact that neurophysiology, as Enc (1983) says, "contains as an essential component a certain abstract level of description of the functional organization of the nervous system" (p. 298), automatically makes it a nonbiological science. Thus, the radical neuron doctrine (RND) as defined is ludicrously strong. G&S's purified definition, excluding all psychological, the-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

oretical, or behavioral terms from neurobiology, allows only theorists who refuse to invoke concepts such as classical conditioning, information, and representation consistently to propose RND. There may be some such theorists among those who deny the utility of current concepts of representation, seeking instead to replace psychology with terms from dynamical systems theory (van Gelder 1995) or even quantum theory (Penrose 1994). G&S could persuasively argue that these attempts to unify cognitive science *directly* with physics, which are compatible with RND, do not have sufficient resources to explain mentality. Surprisingly, though, they are not the targets. Instead, G&S implausibly interpret the Churchlands as supporters of RND. However, neurocomputational models of learning and memory centrally invoke representations (P. S. Churchland & Sejnowski 1992, pp. 141–237). They are pitched "at a decidedly abstract level": The two-pronged framework of transient, occurrent representations, and enduring, dispositional (distributed) representations can in principle be realized in many neurobiological systems (Churchland & Churchland 1996a, pp. 224–30). Indeed recognizably connectionist frameworks of explicit and implicit memory representation were developed by early modern theorists such as Descartes and Hartley, who relied on quite different neurophysiological realizations, in animal spirits and in vibrations, respectively (Sutton 1998; 1999).

G&S have two responses. First, they complain that the Churchlands do nevertheless, in confusion, often defend RND. A more charitable reading would focus less on hyperbolical rhetoric and more on the Churchlands' detailed proposals for specific neurocomputational explanations, where they rely on thoroughly *cognitive* theories, including the opponent process theory of color perception (Churchland & Churchland 1998, pp. 168–72; cf. the use of psychophysical and clinical data on vision in P. S. Churchland & Ramachandran, 1993, and of the cognitive neuropsychology of emotion and decision making in P. S. Churchland 1996).

More substantively, G&S see the only alternative to RND as the weak trivial neuron doctrine, by which "cognitive neuroscience" will explain mentality. They see all versions of the neuron doctrine that allow for relations of integration (rather than exclusion, reduction, or replacement) between psychology and neurobiology as equally "trivial" (sect. 2.2.1, para. 2). G&S's definition of "cognitive neuroscience" is too inclusive. As a label for a "science of minimal commitments," their category includes a "vast family of sciences" that might contribute to an understanding of mentality (Stoljar & Gold 1998, pp. 111, 130). Approaches as diverse as computational neuroscience and cognitive ethology *do* actively seek "to synthesize biology and psychology in order to understand the mind" (sect. 2.1, para. 2), but many others who accept the basic materialism of the trivial neuron doctrine do not pursue this synthesis, and a definition of cognitive neuroscience that includes them is misleading. In, for example, Chomskian linguistics, psychoanalysis, and classical artificial intelligence (AI), many theorists study the *brain* only in the attenuated sense that, say, geologists or ecologists study particles. This is not yet a criticism; it might be (as the analogy makes clear) that direct study of the brain does not aid understanding of some mental phenomena. The Churchlands' targets are not the psychological and linguistic sciences per se, but only certain theories within those sciences. In context, P. M. Churchland's reference to "an alternative to, or potential reduction of" Chomskyan linguistics is clearly not a statement of RND (sect. 1.2, para. 5) but an empirical bet that other (neuro)computational, thoroughly cognitive frameworks will better explain linguistic performance and competence.

G&S see Kandel's account of learning as a mere implementation, rather than a reduction, of psychological theory. This is a controversial, narrow picture of reduction, by which the reducing theory has to be entirely conceptually independent of the reduced theory. However, many philosophers of science hold that reductions can be *partial* (Bickle 1998). In a thoroughgoing discussion of Kandel's work, for example, Kenneth Schaffner (1992, pp. 323–39) argues that reductive connections between psychology and neurobiology need not be simple. He acknowledges that the causal generalizations of theories such as Kandel's are "typically *not* framed in purely biochemical terminology" but instead mix different levels: There is not even a single neurobiological level, as the model of molecular biological processes is integrated into, or "seen as a more detailed expansion of the neural circuit for the gill-siphon reflex." Genuine explanatory reductions will produce "many weblike and bushy connections" across levels, with causal sequences described at many levels of aggregation. The generalizability of biological reductions is limited; some may be specific to the system in question. Thus, not even reductionists impressed by Kandel need claim that this kind of synaptic plasticity explains *all* forms of learning and memory, though Kandel himself seems tempted by RND (1987, p. viii). Reduction, on a range of more liberal views, is "bound to be patchy" (Schaffner 1992, p. 37; cf. P. M. Churchland, 1996, p. 306, on "objective knowledge of a highly idiosyncratic reality").

G&S rely on a sharp distinction between "parasitic" theories, which merely specify implementing mechanisms for independent psychological functions, and genuinely reductive theories (such as the kinetic theory of heat) which render reduced terms ("temperature") explanatorily redundant (sect. 5.3.3, para. 2). In their view, explanations in neurobiology that rely on functional characterizations of the explananda are automatically (nonreductive) mere implementations. However, if a Schaffner-like picture of reduction is correct, this distinction breaks down, and many different relations *between* mere implementation and complete reduction are possible. A modified "argument from unification as reduction" can then go through. G&S's strategy against this argument (sect. 4.2) is to set aside the "enormous literature dealing with reductionism" and then to interpret reductionism in a specific, implausibly strong way, as requiring direct and complete descent to the physical. If this was the only form of reductionism, then reductionists would refute themselves whenever they employ terms other than those of a completed fundamental physics, but it is not. The modified argument from reductive unification encourages close engagement, as exemplified by G&S, with the complex mesh of causal generalizations embedded in specific neurophysiological theories and importantly leaves open the possibility that, in some domains, psychological concepts may be (partially) revised. RND then becomes unnecessary; we get a modified conception of genuine reduction without inevitably dispensing with psychological concepts.

Taking the trivial doctrine seriously: Functionalism, eliminativism, and materialism

Maurizio Tirassa

Centro di Scienza Cognitiva, Università di Torino, 10123 Turin, Italy.
tirassa@psych.unito.it

Abstract: Gold & Stoljar's (G&S's) characterization of the trivial doctrine and of its relationships with the radical one misses some differences that may be crucial. The radical doctrine can be read as a derivative of the computational version of functionalism that provides the backbone of current cognitive science and is fundamentally uninterested in biology: Both doctrines are fundamentally wrong. The synthesis between neurobiology and psychology requires instead that minds be viewed as ontologically primitive, that is, as material properties of functioning bodies. G&S's characterization of the trivial doctrine should therefore be correspondingly modified.

Gold & Stoljar (G&S) contrast two versions of the neuron doctrine, the claim that scientific understanding of the mind will come from neurobiology. In the trivial version, understanding will require the synthesis of neurobiology and psychology; according

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

to G&S, this reading of the neuron doctrine is uncontroversial and widely accepted in cognitive science. The radical version is instead eliminativist: It can be clearly distinguished, at least in general, from the trivial one, and G&S prove it wrong by showing that precisely the psychological concepts it rejects are instead necessary in neurobiology.

This picture, however, misses some differences that may be crucial. The backbone of cognitive science is currently provided by a doctrine, computational functionalism, that has nothing to do with biology, which it views as mere implementation, that is, accidental. Furthermore, this doctrine has close (albeit seldom acknowledged) relationships with eliminativism. G&S's characterization of the trivial doctrine as the assembly of all noneliminativist views of cognition blurs the difference between the biologically inspired and the classically computational views of cognition and should therefore be correspondingly modified.

That the computer metaphor lies at the heart of most contemporary studies of the mind hardly seems debatable: Talk of cognition as computation is commonplace both in the symbolic and in the connectionist literature, albeit with different specifications, and so is the tenet that the architecture of the mind is that of an information-processing system. (Let us set aside the exhausting discussion of precisely what sort of information-processing system the mind is supposed to be.) It is constitutive of this perspective that the mind is an abstract description of the physical machine that happens to "implement" cognition. The doctrine of multiple realizability is the natural, if sometimes disowned, child of this view.

Computational functionalism is generally held to be the opposite of eliminativism; indeed, this seems to be G&S's view as well when they exclude from their characterization of the trivial doctrine only "a certain version of artificial intelligence" (sect. 2.2.1, para. 7). As the mind is stripped of its ontological primitiveness, however, the radical doctrine loses much of its radicalness and unreasonableness: After all, if the mind is only an abstract level of description, why should cognitive scientists not simply do away with it and focus instead on the concrete physical machine? Eliminativism is another natural child of computationalism, though, again, one that is often disowned. (There is no space here to discuss how it relates to the doctrine of multiple realizability.)

Should this reading appear somewhat wicked, reconsider in its light the quotation that G&S make from Higginbotham[1] (1990): "the study of the mind is the study of the brain and nervous system, conducted at some level of abstraction that we would like to clarify" (sect. 2.2.1, para. 4).

As another example, suppose that, thanks to some novel mathematical approach, a yet-to-be-devised nonassociative connectionist network proves capable of satisfactory grammatical parsing: Would this not be a proof, to those who endorse computational functionalism, that the alleged abstract level of description is fundamentally useless and thus count as a match point for the radical doctrine that minds are but the folk postulates of an immature science? (According to the account that the mind is a computational device implemented in the brain, such a hypothetical network must be at least a possibility, if computationalism is not to espouse dualism. The insufficiencies of current connectionism thus cannot be used to do away with this point.)

The only way out of these two related versions of eliminativism (computational functionalism and G&S's radical doctrine) is to acknowledge that minds are ontologically primitive rather than disposable high-level descriptions of what is actually occurring at the physical level. Unless one is willing to be a dualist, this position in turn entails that minds must be conceived of as material properties of functioning brains (or, better still, functioning bodies). As with all natural phenomena, minds can of course be described, but they are not themselves levels of descriptions.

To resume, the situation is much more complex and articulated, and more controversial, than can be captured with the dichotomy between mentalism and eliminativism, unless careful constraints are imposed on what view of the mind counts as noneliminativist. In particular, what G&S call the trivial doctrine comes in two quite different versions that should not be conflated. One version endorses what seems to me (and to G&S) the main tenet of materialist cognitive science, namely, the idea that neurobiology and psychology should proceed toward a nondualist, noneliminativist synthesis. This doctrine will turn out to be correct in one or another of its possible versions and has all the substantive consequences one might desire, concerning in particular the plausibility of computationalism and the role of formal tools and externalized codes in the study of cognition.

Most cognitive scientists, however, seem to endorse a different version of the trivial doctrine, one that builds on the computational version of functionalism that has so pervasively shaped the development of our discipline and while viewing biology as implementation, is fundamentally uninterested in it. The radical doctrine may be viewed as a consequent, albeit somewhat perverted, derivative of this stance, and both are fundamentally mistaken.

NOTE

1. My point is not particularly aimed at Higginbotham's work or at linguistics in general: Analogous statements abound in the literature on the philosophy of cognitive science as well as in textbooks and introductions to the field.

Let us keep our ontology and epistemology separate!

William R. Uttal

Arizona State University, Department of Industrial and Management Systems Engineering, Tempe 85287-5906. **aowru@asu.edu**

Abstract: Gold & Stoljar are right in their thesis but incomplete in not pointing out that there are many other arguments from cognate sciences suggesting that a radical eliminativist neuroreductionism is unlikely to be achieved. The radical neuron doctrine they criticize is only a hoped for dogma that cannot be verified, whereas a constrained monistic materialism (with only partial reductionism) is subject to immediate test by applying such criteria as combinatorial complexity and thermodynamic irreversibility.

Congratulations to Gold & Stoljar (G&S). They have brought clarity and good sense to the absurdities of the hyper-reductive statements of some overly enthusiastic neurobiologists and their attendant philosophers. I fully support G&S's position and offer the following supplementary material to strengthen their case further.

Let us express the key issue using G&S's terminology. First, the "radical neuron doctrine" argument is as follows: Mental processes will ultimately be fully explained by neurobiological data and theories! Insofar as this has not yet happened, it is necessarily an assumption, a dogma, a prejudgment, a hope, and/or an untestable hypothesis. There is no way to verify this conjecture conclusively at the present time. Second, the "trivial neuron doctrine" is as follows: Mind, although assuredly a biological process, cannot be fully explained by neurobiological processes! If it can be shown that there are strong reasons to support this doctrine, then it would be possible to accept it at the present time. I contend that, unlike the case with the radical version, there are solid arguments that make this version into a testable hypothesis.

The following list contains the most compelling epistemological arguments against eliminativism.

1. *Combinatorial explosion:* The number of neurons involved in even the simplest kind of cognitive process is so large that no conceivable model could analyze it at the necessary level of detail from which mentation emerges.

2. *Entropic irreversibility:* There is no way to go from the current state of a complex system such as the mind to its initial conditions, that is, to its neural origins.

3. *Functionally closed systems:* Although the "black box" represented by the brain may be surgically or tomographically

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

opened, because of its complexity it remains functionally closed.

4. *Chaotic apparent randomness:* Although determinist in principle, information about the brain is in a state of functional or apparent randomness. It is impossible either to work up from the neural component level to mind or to work backwards from the mind to the network. The necessary information to do either is simply not available.

5. *Descriptive neutrality:* Mathematics, computational models, and psychophysics are only descriptive and are neutral with regard to the specific internal mechanics or processes of a complex system.

6. *Misinterpreted analogies:* Low-level neural processes are not the equivalents of perception; transmission codes represent stimulus parameters but are not the same as the mechanisms instantiating the veridical mental responses. Furthermore, the responses of single neurons are not the psychoneural equivalents of mental processes.

7. *Cognitive penetration:* The semantic content of a message affects its interpretation; thus, strict neuroreductionism ignores some of the critical high-level influences on mental acts.

8. *Dynamic, adaptive, statistical, and redundant nature of neural states:* The neural net is so unstable, redundant, and adaptive that it is unlikely we will ever be able to define a unique equivalence of neural network state and mental process.

By ignoring these arguments, eliminativists confuse their untestable ontological beliefs with testable epistemological constraints. Furthermore, the eliminativist position confuses many contemporary neuroscientific accomplishments with future goals that are beyond any plausible scientific approach. This is not to say that mental processes are not material manifestations or that they reflect some kind of dualistic or metaphysical reality. Rather, it is to emphasize that there are well established limits and boundaries on the acquisition of knowledge established by other kinds of normal science that affect the extreme neuroreductionist dogma. It is also to say that, although considerable progress has been made in understanding the simplest of neural processes and codes (usually those of transmission and relatively peripheral sensory or motor cortical mechanisms), we have no knowledge of how the activity in complex networks of central neurons become the equivalents of (or are) mental processes.

As impressive as is Kandel's research, G&S's extensive discussion of it is largely irrelevant. As they point out, arguing from functional analogies such as the one between conditioning in humans and the simple kind of learning in *Aplysia* is a typical source of the eliminativist's false optimism. Indeed, Hawkins and Kandel (1984) themselves noted that they "do not provide any data suggesting that higher orders of conditioning must necessarily emerge from the basic cellular mechanisms of more elementary forms of learning" (p. 389). One hopes that this wisdom still holds in their thinking.

Finally, nothing said here should be interpreted to mean that neurobiology has no place in psychology (and certainly not vice versa). Neurophysiology has added enormously to our understanding of how we see and move. The problem is, as G&S have so properly noted and as I have argued more extensively elsewhere (Uttal 1998), a radical neuroreductionism is unlikely to eliminate completely molar psychological theories and concepts.

Synaptic plasticity is complex; neurobiologists are not

Richard M. Vickery

School of Physiology and Pharmacology, University of New South Wales, Sydney, NSW 2052, Australia. **richard.vickery@unsw.edu.au**
www.med.unsw.edu.au/Physiology/school/staff/vickery/welcome.html

Abstract: The complexity of modern neurobiology in even a comparatively restricted area such as use-dependent synaptic plasticity is underestimated by the authors. This leads them to reject a neurobiological model of learning as conceptually parasitic on the psychology of conditioning, on the basis of objections that are shown to be unsustainable. An argument is also advanced that neurobiologists hold an intermediate version of the neuron doctrine rather than a conflated one. In this version, neurobiologists believe that psychology will eventually be underpinned by neurobiology but are agnostic about the extent of upheaval that this will produce in psychology.

The argument from an exemplar in favor of the radical neuron is rejected by Gold & Stoljar (G&S) on the grounds that Kandel's theory of learning in inadequate to explain the importance of timing and stimulus properties in classical conditioning. However, there are sound neurobiological concepts in the literature of synaptic plasticity to explain both of the cited counterexamples.

First, G&S discuss an experiment (sect. 5.3.5) in which a conditioned response develops only when the unconditioned stimulus (US) is not on during periods when the conditioned stimulus (CS) is off. Precise temporal and phase relationships between stimuli are very important in neurobiology, and one explanation of these experimental results is an activity-dependent form of synaptic plasticity called "heterosynaptic long-term depression" (LTD; see Bear & Abraham 1996). LTD is synaptic change that results in a decreased efficacy of synaptic transmission. The heterosynaptic form of LTD occurs in synapses that are silent but are proximate to active synapses. In the cited example, during periods when the CS and US are both on, there will be synaptic activity at both synapses, and there will be potentiation (in Kandel's case, presynaptic facilitation). When the US synapse is active but the CS synapse is silent, then the CS synapse will undergo heterosynaptic depression (consistent with the original paper: Rescorla 1968). Therefore, alternating periods of paired and unpaired activity in the CS and US synapses will cause both potentiation and depression that may cancel each other and lead to no conditioned response. This would be a logical prediction from the known neurobiology and would seem to provide a "low-level mechanical process in which the control over a response is passed from one stimulus to another (sect. 5.3.5) without the need to appeal to a richer level of information.

The second cited experiment (sect. 5.3.5, para. 5) demonstrates that not all stimuli can be conditioned with equal ease, and the authors conclude that "a mental representation of a relation governing the stimuli thus has a differential effect on the course of learning." However, this differential effect on learning may simply be due to the structural relationships of population of neurons, in particular, the extent of overlap of axonal and dendritic arbors. Most contact between neurons in the mammalian cortex is through more than a single synaptic bouton, and one axon termination may take many synaptic contacts with the dendrites of a single neuron. The efficacy of the axon in firing the target neuron is related to the number of contacts, their proximity to the target neuron soma, and the specific properties of each synaptic contact. One of several possible neurobiological explanations for stimulus preference in the association experiment is that plasticity requires spatial matching of input synapses on the dendrites of the target neuron (see White et al. 1990). Some types of synaptic plasticity may affect only CS synapses on the same dendritic branch or spine as the synapses of the US input (Denk et al. 1996). Insofar as neuronal inputs are often sorted or ordered in some way, it is possible that the synapses made by the CS "triangle" input are too remote from the US "red square" synapses for there to be any interaction

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

that leads to plasticity, whereas synapses from the CS "square" input may well be located close to the US "red square" synapses. It seems that not only the psychological complexity of simple models of learning has been underestimated but also the neurobiological complexity.

The argument introduced in section 5.3.1 about color opponency is also unconvincing. G&S state that neurobiology provides a description of cell populations with center/surround structure but that it cannot model the function of these cells. This seems false; the population of cells includes only red/green and blue/yellow pairings. The inference a neurobiologist draws from this is that red and green cannot be experienced in the same location at the same time because they are mutually inhibitory. The same applies to yellow and blue. Red can coexist with yellow or blue, and green also with yellow or blue. The neurobiology leads to the same conclusion about function that the psychological theory does, but by an independent route.

A restatement of the neuron doctrine that I believe more accurately reflects the views of neurobiologists might be:

> Neurobiology will come to underpin psychology. This will likely lead to substantial revolution and revision of existing psychological theories. Psychological theories not underpinned by neurobiology will be discarded in favor of those that are.

At present, neurobiology is generally reducible, through biology and biochemistry, to any desired level of explanation, although this is not possible for psychology. This version of the neuron doctrine holds that psychology will ultimately be reducible to neurobiology, and so further down to any other explanatory level, and that, because there are currently few points of contact, this is likely to entail a major revision of psychology. G&S make clear in note 30 that they are discussing reduction in principle, and that "what determines the form of the successful theory is where the best explanation is to be found" (sect. 5.3.6), although I would qualify this to include predictive power as well. Psychological concepts will be used to describe the ensemble activity of neurons but will always be directly understandable in terms of the activity of all the individual neurons in the ensemble. This version seems to give the best reading of G&S's quotes by neuroscientists such as Barlow (sect. 1.1, para. 15) and Churchland and Sejnowski (sect. 1.2, para. 6), and it is neither trivial nor unsubstantive.

Neuronal connectivity, regional differentiation, and brain damage in humans

Dahlia W. Zaidel

Department of Psychology, University of California at Los Angeles, Los Angeles, CA 90095-1653. **dahliaz@ucla.edu**

Abstract: When circumscribed brain regions are damaged in humans, highly specific impairments in language, memory, problem solving, and cognition are observed. Neurosurgery such as "split brain" or hemispherectomy, for example, has shown that encompassing regions, the left and right cerebral hemispheres, each control human behavior in unique ways. Observations stretching over 100 years of patients with unilateral focal brain damage have revealed, without the theoretical benefits of "cognitive neuroscience" or "cognitive psychology," that human behavior is indeed controlled by the brain and its neurons.

The arguments presented by Gold & Stoljar (G&S) have narrow definitions of the relationship between neurons and the mind, particularly when they apply to the human brain. My argument emphasizes what we know about the organization of the human mind in the brain from studying neurological and neurosurgical patients with focal brain damage (DeRenzi et al. 1968), commissurotomy (Bogen 1992; Zaidel 1990; Zaidel & Sperry 1974), and hemispherectomy, and without the benefits of what is today called *cognitive neuroscience*. The behavioral consequences of neuronal connectivity disruptions have been amply verified postmortem and with brain imaging techniques (Geschwind & Galaburda 1984).

First, not all brains are the same, even if they all have neuronal cells. *Aplysia,* rats, birds, monkeys, chimps, and humans do not have the same brains, nor do all species have the same sensory organs to process incoming stimuli. What matters is how neuronal populations have assembled into regions in each type of brain. Brain regions are defined by characteristic neuronal cell size, shape, orientation, axons, dendrites, and neurochemical and physiological processes (Zaidel et al. 1997). In humans the regions exert specialized control over behavior to an extent not seen in other animals, whether mammals or not.

Second, evolutionary adaptive changes have put constraints on the relationship between brain and mind. For example, the hallmark of human cognition is hemispheric specialization (Sperry 1974). The lateralization of speech and most components of language to the left cerebral hemisphere and of topographical knowledge and visuospatial and facial perception to the right hemisphere is unlike anything seen in animals in scope and extent. As the human brain evolved, the major interhemispheric tract of connecting fibers, the corpus callosum, grew to a size larger than in any other mammal. Similarly, as the brain evolved, the hippocampal commissure became smaller, suggesting that, rather than the abundant direct communication between the two hippocampi seen in rats, cats, or monkeys, in humans each hippocampus communicates with the ipsilateral neocortex (Amaral et al. 1984; Rosene & Van Hoesen 1987). Such an arrangement could explain why unilateral hippocampal damage in humans results in memory impairment consistent with the cognitive deficits following neocortical damage on the same side (Zaidel et al. 1994), whereas, with experimental animals, rarely if ever does memory impairment follow unilateral hippocampal damage, with bilateral damage required to produce the impairment. The encompassing region represented by each cerebral hemisphere thus controls different components of the human mind.

Third, observations on the consequences of brain damage do not require for their interpretation theories of cognitive psychology, cognitive neuroscience, or just plain psychology. When a right-handed person suffers from a stroke or a tumor affecting his left hemisphere, particularly the lower third frontal convolution (Broca's area), aphasia, a severe inability to communicate linguistically, emerges. This is simply obvious. The ancient Greeks had already observed this relationship between language and the left side of the brain.

Fourth, not all scientific pursuits of the mind have equal success in uncovering the relationship between neurons and the mind. Neuropsychology and neurology are neurodisciplines that have provided insights on mind and brain through the understanding that focal brain damage in humans fractionates the components of the mind, which, in turn, can be subjected to systematic analysis. Theories of the mind-in-the-brain gleaned by researchers in these disciplines well precede the relatively recent theories of cognitive neuroscience or cognitive science. The building blocks of the mind can be revealed by observing the alterations in neuronal connectivity.

Answers to questions such as "how does human language occur in the first place?" are elusive. The left hemisphere is critically involved; we know that much. How neurons produce language is something that we simply do not yet know (Scheibel 1984). Neuropsychologists and neurologists in collaboration with cognitive neuroscientists and scientists from neuroanatomy, immunohistochemistry, cellular biology, and experimental neuropathology will most likely discover the answer. Interdisciplinary collaborations have provided evidence for a strong relationship between neuronal density in the hippocampus, for example, and memory (Zaidel & Esiri 1996), particularly verbal memory (Rausch & Babb 1993; Sass et al. 1990), and for explicit versus implicit memory (Zaidel et al. 1998). At the same time, associations between mor-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

phological or immunohistochemical features of neurons and components of the mind are sorely missing. In any case, neuronal connectivity surely plays a critical role in producing the mind in the brain.

Playing with words, working with concepts, testing ideas

J. M. Zanker

Center for Visual Sciences, Research School of Biological Sciences, Australian National University, Canberra, ACT 2601, Australia.
johannes.zanker@anu.edu.au
cvs.anu.edu.au/johannes/johannes.html

Abstract: Gold & Stoljar's attempt to disentangle the body-mind problem in time for the end of the decade of the brain deserves praise for its diligence and courage in moving onto the treacherous ground of interdisciplinary discourse. In making their point, they should not have stopped half-way: a more clearly defined experimental paradigm seems necessary to solve this exciting and substantial problem.

Gold & Stoljar (G&S) attempt to assess the validity and explanatory power of a future neuroscientific theory of mind. In doing so, they suggest yet another definition of the term "neuron doctrine." Some decades ago, this term was introduced for Cajal's fundamental claim that separable nerve cells are the basic unit of brain function, instead of a continuous network as postulated by the recticular theory (Shepherd 1991). Since the seminal paper by Barlow (1972), the same label has also been used to distinguish the concept of encoding complex objects in a single neuron from that of representation by assemblies of cortical cells. Now we are confronted with G&S's redefinition of this term to mean that "a successful theory of the mind will be a solely neuroscientific theory." The final aim of this both naive and provocative claim is to demonstrate that neuroscience will provide definitive answers to fundamental questions about the human condition that philosophy seems to have failed to provide for more than two millenia.

The central problem with the G&S's position, exemplified by this change in the meaning of "neuron doctrine," is that crucial arguments eventually end up as difficult semantic problems. For example, the essential aspects of the mind that the authors expect to be explained by neuroscience are paraphrased vaguely as "mental function" or "psychological phenomena," without any further specification. Even worse, the meaning of apparently simple but critical terms such as "neuroscience" remains a matter of dispute. G&S rather deliberately draw a line between the core disciplines of neurophysiology, neuroanatomy, and neurochemistry, which they call "biological" neuroscience and other, psychologically oriented, disciplines constituting a "cognitive" neuroscience. This separation is crucial to their distinction between an undisputed but "trivial" neuron doctrine and a "radical" but questionable one. But what about the rapidly growing and evolving disciplines such as neuroethology and psychophysics, which exactly and quantitatively pin behavior or perception down to their neuronal mechanisms (see, e.g., Spillmann & Werner 1990)? They are excluded from the inner circle of "psychology-free" paradigms, although they could be regarded as frontrunners in the quest for the holy grail of neuroscience as set up by G&S.

G&S identify an important, if not the ultimate, question of modern neuroscience and discuss it with great philosophical rigor, well informed by the neuroscientific literature. They successfully object to a technically minded approach, such as that advocated by Zeki (1993), who praises the collection of scientific facts over "endless and fruitless philosophical discussions about the meanings of words," and they successfully oppose the defiant belief of authors such as Crick (1994), who writes a whole book on "consciousness" without accepting any need to define the word. However, although they make progress in their attempt to identify concepts, to disentangle hidden assumptions and to clarify crucial terms, G&S eventually get caught in a culture of misunderstanding that accompanies the discourse between neuroscience and philosophy. Their logical treatments of the naturalism/materialism argument and of unification, and even their evaluation of a particular exemplar, crucially rely on a semantic issue: What is neurobiology? In consequence, their rejection of the neuron doctrine is based on terminological usage.

Because all of G&S's arguments eventually fall back on the question of which disciplines will be embraced by neurobiology strictu sensu, their discussion of an exemplar seems most promising. Surprisingly, however, they discuss the cellular learning paradigm in *Aplysia*, eventually rejecting its claim to describe a "pure" neurobiological basis of a mental phenomenon, because of contamination by the psychological theory of classical conditioning. This argument leaves the reader puzzled as to how a psychological narrative can be totally avoided. After all, this is the "mental phenomenon" that has to be explained? One also wonders why this particular example was chosen. Why should the explanatory power of neuroscience be tested against the behavior of a lazy sea slug? Would it not be better to talk about neurons performing logical operations, physiological evidence for language processing in the brain, recording of activity believed to be related to a theory of the mind, or neurophysiological explanations of visual illusions offered by psychophysics?

In the twilight zone between science and the humanities, it is important to reach methodological agreement, and, with their argument from exemplars G&S offer neuroscientists an empirical answer. But how far does this option take us? What kind of exemplar would pass their scrutiny? Such an example certainly cannot be a complete description of the mechanism, in depth (down to single molecules) and in scope (embracing all the elements involved). Gierer (1983) doubts whether such an approach is possible. In any case, a complete description does not yield much explanatory power, because a mirror image does not imply understanding (the translation of a Shakespeare sonnet into Japanese does not provide any clue to understanding its meaning).

So, what will be the criteria for a valid example? What would be acceptable as "independent explanation"? How many exemplars and what detail would be required to satisfy G&S? A productive outcome of the discourse between neuroscientists and philosophers (if they could achieve agreement among themselves) would be the formulation of criteria for a kind of neuro-mind Turing test to guide experimental and theoretical efforts by neuroscience to tackle the body–mind problem. Without such a set of explicit criteria, G&S's argument remains in a vacuum, leaving open the prospect of a revival of a dualist position, as proposed by Chalmers (1996).

Difficulties in interpretation associated with substitution failure

Eric Zarahn

Department of Neurology, University of Pennsylvania, Philadelphia, PA 19104. **ericz@mail.med.upenn.edu** **cortex.med.upenn.edu/~zarahn/**

Abstract: In one of their arguments against the radical neuron doctrine, Gold & Stoljar (G&S) use the idea that, in certain situations, equivalent terms may not be substitutable into statements that regard properties of the objects to which the terms refer. This device allows G&S to refute the necessity of the conclusion that "the science of the mind equals the science of the brain" even though they take as a premise that the mind equals the brain. I argue, however, that this practice leaves the meaning of the "science of the mind" and the "science of the brain" indeterminate.

If A and B represent two propositions, then "A is equivalent to B" means that A is true if and only if B is true. In basic logic, a conse-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

quence of the equivalence of two propositions, A and B, is that, wherever A is found as part of another proposition C, it can be replaced by B without changing the truth of C. The qualification is added because there are other cases in which this ability to substitute may be argued to fail. Gold & Stoljar (G&S) claim that such a case exists with the terms "mind" and "science of the mind" (as well as with "brain" and "science of the brain"). The statement that "mental phenomena are identical to neural phenomena" is used by G&S as "a version of" (sect. 3, para. 2) the definition of materialism. However, the authors state (in n. 5) that mind equals brain need not imply that "the science of the mind equals the science of the brain." G&S's argument for this claim may involve treating "science of the mind" (or "science of the brain") as a belief that need not adhere to constraints that we apply to "mind" (or "brain").

G&S use the identity/equivalence of mind and brain quite precisely (for example, "we [the authors] agree also that mental phenomena are identical to neural phenomena"; sect. 3.2, para. 1). There are, as one might imagine, other versions of materialism besides equivalence. An example of such a relationship is that "the mind supervenes on the brain"; this would mean that any change in the mind would be accompanied by a change in the brain (which is not necessarily the same as equivalence). Though the authors acknowledge these other versions, they state (in n. 4) that the equivalence premise "will do."

Is the particular choice of premise of G&S that "mind equals brain" (as opposed to, say, "mind supervenes on brain") really so unimportant to their arguments about both the radical and the trivial neuron doctrines (i.e., the relationship between "the science of the mind" and "the science of the brain")? Do they actually deduce any conclusions from this premise (i.e., is it used as a premise, or is it simply an isolated statement)? I will argue that G&S have made any premises about the "mind" and the "brain" inconsequential to conclusions regarding the "science of the mind" and the "science of the brain." This is because their use of substitution failure in conjunction with their lack of clarification of the precise relationship between the "mind" and the "science of the mind" and the "brain" and the "science of the brain."

First, let us examine, as naively as we can, the consequences of G&S's premise that the brain is equivalent to mind. Would it not follow that every property of the "brain" would have to apply to the "mind" as well (and vice versa)? If so, for example, "the science of the mind" would have to equal "the science of the brain."

In section 3.1, however, G&S appeal to the idea that the equivalence of the "science of the mind" and the "science of the brain" does not necessarily follow from the equivalence of "mind" and "brain." They use this substitution failure as the primary method of refutation of an argument supporting the radical neuron doctrine, which they considered in section 3. However, given this appeal to substitution failure, it seems that the authors have made something of a superfluous statement in "the mind is equivalent to the brain" because, owing to the failure of substitution of the terms "mind" and "brain" into the statement "the science of the mind is equivalent to the science of the brain," their premise seems to have no bearing whatsoever on the relationship between the "science of the mind" and the "science of the brain." More generally, it would follow that any property of the "brain" need have no bearing on the "science of the brain" (and likewise any property of the "mind" need have no bearing on the "science of the mind"). This does not seem to be a mistake or oversight on the part of the authors but, rather, the intended method of this particular refutation of the necessity of the radical neuron doctrine. This particular device, however has dialectical reverberations. In this appeal to the failure of substitution (and in the absence of further elaboration on the presumed relationship between "X" and the "science of X"), G&S sever the formal relationship between the terms "mind" and "science of the mind" (as well as that between "brain" and "science of the brain"). From this perspective, any further inquiry into the relationship between "the science of the mind" and the "science of the brain" seems uninteresting, or at least obscure.

Regardless of how it was actually used (or not used) by G&S, the premise of equivalence itself has interesting consequences if we keep the premise of mind–brain equivalence and do not invoke failure of substitution. This would imply that "the science of the brain is equivalent to the science of the mind," and hence that the "mind" could be explained completely by G&S's "biological neuroscience" (sect. 2.1, para. 1). However, it would not imply the radical neuron doctrine (as incorrectly concluded sect. 1, para. 3) because "the science of the mind" would conversely be itself capable of complete explanation of the "brain." That is, equivalence is symmetric, and "the science of the brain is equivalent to the science of the mind" gives no favor to "science of the brain" over "science of the mind." We might also wonder if the naive premise of equality seems valid. On first consideration, the answer would have to be no. This is because the "brain" has properties that would not agree with most definitions of the "mind" (e.g., the statement "the mind is bathed in cerebrospinal fluid").

ACKNOWLEDGMENT

I thank Jennifer Saul for sharing her extremely edifying ideas regarding substitution of coreferents.

Authors' Response

Interpreting neuroscience and explaining the mind

Ian Gold[a] and Daniel Stoljar[b]

[a]Institute of Advanced Studies, Australian National University, Canberra ACT 0200, Australia, and Department of Ophthalmology, Royal Victoria Hospital, Montreal, Quebec, Canada H3A 1A1; [b]Department of Philosophy and Institute of Cognitive Science, University of Colorado, Boulder, CO 80309, and Institute of Advanced Studies, Australian National University, Canberra ACT 0200, Australia. **{iangold; dstoljar}@coombs.anu.edu.au ian@vision.mcgill.ca stoljar@colorado.edu www.coombs.anu.edu.au/Depts/RSSS/Philosophy/People/{IanGold; Stoljar}.html**

Abstract: Although a wide variety of questions were raised about different aspects of the target article, most of them fall into one of five categories each of which deals with a general question. These questions are (1) Is the radical neuron doctrine really radical? (2) Is the trivial neuron doctrine really trivial? (3) Were we sufficiently critical of the radical neuron doctrine? (4) Is there a distinction to be drawn at all between the two doctrines? and (5) How does our argument bear on related issues in the ontology of mind? Our replies to the objections and observations presented are organized around these five questions.

Central to the target article is an analysis of the theory of elementary learning in *Aplysia californica* developed by Eric Kandel and his coworkers. According to the analysis, Kandel's theory has two parts. The first part provides what Robert Cummins (1983) calls a *property theory* and explains what elementary learning is. As we argued, this part of the theory relies on the psychology of classical conditioning. The second part provides what Cummins calls an *instantiation theory* and describes how the psychology of classical conditioning can be implemented in *Aplysia* neurons. If this analysis is correct, then to those who claim that neuroscience will ultimately explain mental phenomena one can pose an important question, namely, "What do you mean by 'neuroscience'?" If you mean neuroscience as typ-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

ified by the work of Kandel and coworkers, the claim that neuroscience will ultimately explain mental phenomena can only be interpreted as the trivial neuron doctrine (TND) – trivial in the sense that it is committed only to an explanation of the mind by some collection of relevant sciences. In the case of Kandel's theory, both neurobiology and the psychology of classical conditioning are relevant. If you mean something more stringent than Kandel's theory, however, such as the view we call the radical neuron doctrine (RND), according to which neurobiology alone will explain the mind, then your view is not supported by our best current science. Nor is it supported by general considerations of philosophy and the history of science. The best evidence we now have, therefore, supports only the weak claim that the successful theory of the mind will be an eclectic one. The evidence may change, but at the moment, the rational view is an agnosticism about the possibility that neurobiology alone will explain the mind. The significance of this conclusion is that many in the field believe, or seem to believe, that the opposite is true.

The commentaries on the target article provide many different points of view about our interpretation and evaluation of the neuron doctrine. Some argue that we are obviously right, some that we are obviously wrong, and some that we did not go far enough. We thank all the commentators for their observations and for the challenges they have presented to our position. We have tried to address all of the main criticisms they have made, but, for reasons of space, we have regrettably not been able to address every point. In addition, although many commentators expressed support for specific aspects of our argument, our remarks naturally focus on areas of disagreement or apparent disagreement.

We divide our replies into five main themes: (1) whether the RND is really radical; (2) whether the TND is really trivial; (3) whether we went far enough in our criticism of the RND; (4) whether there is a distinction to be drawn at all between the two doctrines; and (5) what relations our position bears to the ontology of mind.

R1. Is the radical neuron doctrine really radical?

Many commentators suggest that we are wrong to think that the RND is radical, or at least that it is radical in some objectionable way. There are a number of different objections which we discuss in turn.

R1.1. Relevant concepts. Many commentators raise a question about the concepts to which a neurobiological theory of the mind can properly appeal. **Sutton** takes our characterization of the RND to include the claim that neurobiology is prohibited from making use of abstract functional concepts in order to explain behavior. According to this view, any theory likely to produce an explanation of behavior would immediately count as non-neurobiological and would be, as Sutton says, a ludicrously strong position.

Sutton is right that this would be much too strong. Our claim, however, is not that neurobiologists are prohibited from appealing to functional descriptions but rather that when they do so it is often *psychology* that provides these descriptions. The example of Kandel bears this out, but we also think that other examples, such as that of long-term potentiation (see Stoljar & Gold 1998), are best analyzed in a similar fashion. Sutton wants to defend a position that is neither the RND nor the TND as we describe them, but in order to do so, he needs to do more than claim that neurobiology can help itself to notions of structure and function. He needs to show, rather, that we are mistaken in our approach to Kandel's theory by arguing that it is either not a successful bit of neuroscience, or else that it is not a typical bit of neuroscience.

Lau asks whether we intended to include computational or representational notions in biological neuroscience. If not, then the RND is obviously false because no theory of the mind that fails to include computation or information can be successful; but if so, then our account of Kandel's theory is weakened. We argued that the theory cannot plausibly be interpreted as a reducing theory of classical conditioning because the neurobiological concepts Kandel employs – synaptic plasticity in particular – cannot reconstruct central features of classical conditioning such as the notion of an animal's representing relations among stimuli. However, Lau argues, if representational notions are part of the arsenal of the neurobiologist, then, for all we know, a future extension of Kandel's account will be able to reduce the psychology of classical conditioning.

Lau is quite right that a future version of neurobiology (perhaps including computational and representational notions) might provide a reducing theory for classical conditioning. But he is mistaken in thinking that this would affect our analysis of Kandel's theory. The crucial fact about Kandel's account, as we see it, is not that the psychological theory it needs is computational or representational but that it is a piece of psychology that cannot at the moment be discarded. It is reasonable to envisage a future theory somewhat like Kandel's that would also provide a reducing theory. But it is not to envisage Kandel's theory.

Setting aside **Lau**'s specific suggestion concerning Kandel, his question about representational and computational notions is important in its own right. It is uncontroversial that neurobiology can appeal to abstract functional notions, and, in some uses of the term, these might reasonably be called computational. Neurobiology also includes notions such as that of a receptive field, which might reasonably be called representational. The mere fact, therefore, that a theory incorporates representational and computational notions does not prevent it from being neurobiological. The crucial question for us is whether the representational and computational notions have a more natural home in psychology or in neurobiology, and that question can only be answered on a case by case basis. The crucial fact about the neurobiology of elementary learning is that it is not reducible to neurobiology *as it currently is*, and this is all that is required for us to reject the argument from exemplars. Perhaps Lau is suggesting that neurobiology might develop new representational and computational notions that are totally unlike its current ones. There are various ways that this might come about (we describe them in sect. 4 of the target article), but none of them supports the RND.

Both **Sutton** and **Lau** are interested in the possibility that neurobiology *itself* might develop abstract accounts of cognitive function. In contrast, **Blumenthal & Schirillo** suggest that evolution might provide such accounts, and that these would allow neurobiology to do without psychology.

So far as we can see, evolutionary considerations cut across the distinction we want to focus on. It is certainly possible that evolution might provide an abstract account of

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

cognitive function. Until such an account is provided, however, we cannot tell whether it will make neurobiological accounts of mental function easier to develop or not. For all we know, evolutionary accounts of mental life will mesh better with psychology than with neurobiology. After all, there are a number of central evolutionary notions – adaptation, fitness, and environment, to name some of the familiar ones – that are no more at home in neurobiology than in psychology. In fact, evolution has been applied (e.g., by Barkow et al. 1992) to explain psychological structure and function. For this reason, although evolutionary considerations *might* support an RND, we have as yet no evidence that they do.

R1.2. Linguistic issues. According to **Hardcastle,** we have made the RND radical only by engaging in armchair neuroscience. She seems to suggest that the question of which terms belong to which science is not significant, or, at any rate, not one to be answered by philosophers. The fact that Kandel's theory appears to explain what looks like psychological phenomena is really just an illusion of terminology. One might equally be misled into thinking that since the immune system "learns" to create antibodies, theories of this process are psychological theories. No doctrine of interest, certainly no "radical" doctrine, should be decided on terminology.

If **Hardcastle** is right, talk of learning in *Aplysia*, like talk of learning in the immune system, is only metaphorical. However, as she herself emphasizes, one of the interesting aspects of Kandel's work is that it may generalize and tell us something universal about the nature of memory itself, in particular about the distinction between long- and short-term memory. But it is hard to see how this could be true if we interpret attributions of memory to sea slugs as metaphorical as we do in the case of the immune system.

Zanker's commentary raises a related point. He argues that we have attempted to answer substantive scientific questions by semantic means. As we understand him, Zanker's claim is that we have artificially restricted the domain of the term "neurobiology" and have thereby ignored other areas of neuroscientific research – he mentions neuroethology and psychophysics – in which neural explanations of psychological phenomena may be available. We have, therefore, made it extremely difficult, if not impossible, for the RND to come out true. Zanker further claims that this restriction leads us to choose the inappropriate exemplar of Kandel's theory against which to test our claims. Here, however, Zanker's claim is not that we have applied semantics where science is necessary, but that the RND can only fail to be supported by the exemplar because every psychological phenomenon requires a psychological description to identify it at all. Isn't a solely neural account of the exemplar therefore impossible from the start?

We did note in the target article (in n. 17) that less radical neuron doctrines could be produced by increasing the number of branches of neuroscience included in neurobiology. Whether or not neuroethology or psychophysics could be added to neurobiology to produce a neuron doctrine that is interesting is a question we cannot pursue here, but whatever the answer, it does not seem to us to be one that would be determined by linguistic fiat. The borders of neurobiology may be fuzzy, but they are not stipulative. It is possible that there are exemplars that would lend more support to the RND, and we are prepared to consider them on their merits and on substantive scientific grounds.

In any case, even if **Zanker** is right that we have been overly restrictive in our definition of neurobiology, it is hard to see how that criticism supports his claim that we have sneaked psychology in by the back door in our discussion of classical conditioning. Our claim that Kandel's theory incorporates psychological notions is not based on the idea that any description of a psychological phenomenon must appeal to psychological concepts. This would indeed be to establish our position by theft rather than by honest toil. Our claim is that Kandel's theory appeals to substantive psychological notions that cannot be eliminated without significantly limiting the explanatory power of the theory. More important, the psychological theory of classical conditioning is the sort of psychology that seems unambiguously excluded from an expanded neurobiology.

We conclude with a logical point raised by **Zarahn,** who suggests that we overestimate the radicalness of the RND by not considering in detail the status of the inference from "the mind is the brain" to "the science of the mind is the science of the brain." Zarahn takes up a note in which we point out that "the science of" is an intensional functor in the following sense: from "A = B" it does not follow that "the science of A = the science of B."

We are not sure whether **Zarahn** objects to our general claim about "the science of" or whether his concern is the particular case of mind–brain identity. It is unlikely that his concern is with the former because, without doubt, failure of substitution does occur in some cases: for example, planets are identical to clumps of atoms, but the science of planets is not the science of clumps of atoms. If, however, his concern is with the particular case of mind–brain identity, it is hard to see how the inference from "mind = brain" to "science of mind = science of brain" poses a problem for our argument. The concern of the target article is precisely with the question of what meaning should be given to the phrase "science of the brain" when considering its role in the explanation of the mind. Our claim in part is that there *is* a sense in which the science of the mind is the science of the brain, namely, where "science of the brain" is interpreted as cognitive neuroscience. However, one comes down to the question of substitution – even if one supposes that the inference Zarahn discusses is a good one – the question of the neuron doctrine will have to be evaluated on its merits.

R1.3. The resources of neurobiology. The comments by **Hameroff, Zaidel,** and **Vickery** take an empirical approach to the question of whether the radical doctrine is really radical. According to these commentators, we have underestimated the complexity of neurobiology, and this affects our conclusions.

Hameroff suggests that the standard picture of the neuron is simplistic in a way that hides the possibilities for explaining psychological phenomena at the subneural level. A doctrine that makes full use of the complexity of the subneural function would not seem as radical as we take the neuron doctrine to be.

Hameroff's suggestion is an interesting one and deserves to be taken seriously; he is certainly right that the views we discuss tend to neglect the complexity of the neuron. Of course, Kandel's own work is not such a case be-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

cause parts of his theory of learning that we did not discuss refer to gene expression. Hameroff is also right that should neurobiological explanation begin to make use of the concepts of intraneural function to explain psychological phenomena, then we would have to reconsider the plausibility of the RND. We have not argued that the RND is false but that it is currently unsupported by our best science, and scientific progress might require a change in our position. It is not clear, however, that our analysis would *have* to be changed even if future neuroscience begins to exploit intraneural function. The mere fact that one appeals to intraneural complexity in explaining some cognitive phenomenon does not in itself tell us anything about the structure of that explanation. It may continue to be an implementation account rather than a reduction.

Zaidel focuses on neural systems-level explanations of psychological phenomena rather than intraneural level explanations. She argues that neuropsychological research offers an example of psychology-independent investigation of the mind and constitutes an area of the theory of the mind in which the RND is true. We certainly agree that the study of focal (or other) brain damage offers important insights into mental function, in particular, as Zaidel notes, by fractionating behavior. But this is not the RND in practice because, at its best, neuropsychological investigation reveals previously unknown psychological functions. When we discover, for example, that patients with Alzheimer's disease can read fluently without comprehension (Warrington 1975), we are learning something about the function of the brain but not about the function of neurons or neural systems. The dissociation between fluent reading and linguistic comprehension is a cognitive one whose underlying neural implementation is no better understood as a result. In fractionating behavior, neuropsychology contributes to psychology more than to neurobiology.

Vickery argues that we have underestimated the power of neurobiological concepts to account for classical conditioning. He offers neurobiological explanations for the conditioning phenomena we refer to in section 5.3.5 in the target article in order to argue that Kandel's theory could not successfully reduce classical conditioning. He first discusses heterosynaptic long-term depression (LTD) to explain the differential conditioning results obtained when the conditioned and unconditioned stimuli only occur together and when the conditioned stimulus also occurs without co-occurrence of the unconditioned stimulus. He then hypothesizes that possible differences in spatial structure at the level of synapses might account for the differential learning effects in the second-order conditioning experiment we discuss.

We are grateful for **Vickery**'s suggestions, particularly regarding LTD, because he is quite right that there is considerable variety in the mechanisms of potentiation and depression, and there may indeed be other neurobiological resources for explaining learning behavior than the ones we mention. There are two points to be made about his suggestions, however. First, it does not seem necessary to appeal to LTD in order to offer a mechanism for the first result because the failure of the conditioned and unconditioned circuits to be co-active in all cases might be sufficient to explain Rescorla's results. Indeed, one would expect LTD to lead to an increased difficulty in conditioning, whereas Rescorla's (1988) results merely indicate a *failure* of conditioning. Second, describing the correct neurobiological mechanism is not really central to our point. Our argument about the first of the conditioning results is meant only to call into question the adequacy of the received view about conditioning as a transfer of response from unconditioned to conditioned stimulus. If "mere transfer" is not an adequate characterization of conditioning, then Kandel's account (which seems to translate the notion of transfer into neurobiological terms) cannot be used to support the RND despite its contribution to understanding classical conditioning. Even if Vickery is right that Kandel's theory can account for the first of the conditioning experiments, that experiment may nonetheless expose the limitations of the neurobiological resources available to explain conditioning. His suggestion about mechanism may be correct without undermining our claim about the relation between the neurobiology and the psychology of classical conditioning.

Similarly, if the notion of stimulus relations raised in our discussion of the second experiment cannot be captured by the traditional conception of conditioning, then Kandel's theory, which seems to be best suited to the traditional conception, must be described as implementation rather than reduction. **Vickery**'s proposal, as far as we can tell, is that spatial matching is required for synaptic plasticity and acts as a sort of filter: in the "similar" case, the synapses of the conditioned and unconditioned circuit are sufficiently close to undergo facilitation, whereas in the "dissimilar case" they are not. We concede that were an explanation of this sort to be verified, it would go some way toward offering a richer neurobiological story about conditioning. However, would that story reduce the psychological story or implement it? Surely, if some notion of the representation of relations among stimuli is part of the psychological story, then spatial matching by itself would not serve because not all relations are spatial relations. Once again, it is important to emphasize that we are not calling into question the idea that Vickery's proposed account (or some similar one) might be correct. Our claim is only that those mechanisms do not take over the conceptual work done by the psychological theory.

Vickery's point about our account of color opponency is relevant here. He claims that the examination of opponent neurons would produce the same discovery as the psychological theory, namely, that one cannot see a reddish-green or a bluish-yellow. Perhaps, but opponent theory is much more than that observation, and there is as yet no neurobiological story about opponent cells that makes the psychological story otiose. Similarly, Vickery does not offer a set of neurobiological concepts that could replace the psychological ones that explain conditioning. Of course, Vickery may be on to something. There may be some such story waiting to be articulated, and we are prepared to revise our position should it become available.

R1.4. Objectionable consequences. Setting aside the question of the interpretation of the RND, a number of commentators claimed that our interpretation does not entail the consequence we draw. **Byrne & Hilbert** distinguish two radical doctrines and claim that we conflate them to our detriment. The weak doctrine holds that the mind will be explained by neurobiology; the strong doctrine holds that *only* neurobiology will explain the mind. Byrne & Hilbert's claim is that the mere existence of a low-level theory that explains a set of phenomena does not ipso facto entail that the higher-level theory might not also explain the phenomena. For this reason, the future success of neurobiology does not entail

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

that the psychological sciences are place-holder sciences, and neither is this a consequence defended by the Churchlands (e.g., P. S. Churchland 1997). Byrne & Hilbert consider the argument from unification as an instance of our error, but the point is a general one. According to Byrne & Hilbert, the Churchlands hold the weak doctrine, but our critique of the RND assumes that the strong doctrine is true, and for this reason, our argument misses its target.

There are two objections to **Byrne & Hilbert**'s suggestion. The first is that explanations in science tend to exclude each other. Let us suppose with Byrne & Hilbert that there is (or could be) a neurobiological theory of language that is genuinely successful. Such a theory would either be better than traditional linguistics or it would not. (It could be exactly as good, but such cases are the exception rather than the rule in science.) If it were worse, it would not replace the psychological theory and neither the weak nor the strong version of the RND would be true. If it were better, what explanatory significance should we attach to the fact that there is a somewhat less successful psychological theory also available? It is not clear why we should attach any significance to that fact at all. Thus, even if the RND is ambiguous between a weak and a strong doctrine, this does not alter our argument concerning the RND or our analysis of the argument from unification.

The second objection is that although **Byrne & Hilbert** are right that one can draw a distinction between a strong view that *only* neurobiology will explain the mind and a weak view that *at least* neurobiology will explain the mind, it is less clear how this distinction will be of use in practice. The reason is that, as Byrne & Hilbert point out, even according to the weak view, psychology must be interpreted as reducing to neurobiology. Hence, the strong view that only neurobiology will explain the mind does not exclude the claim that psychology will *also* explain the mind since a reduced psychology *just is* a part of neurobiology. (Compare: to say that only philosophers will come to the party does not rule out the possibility that Jones will come, so long as Jones is a philosopher.) The distinction drawn by Byrne and Hilbert and their critique of the argument from unification is thus in danger of collapse. Of course, this leaves open the question of whether psychology *does* reduce to neurobiology – but that is precisely one of the questions that our paper addresses.

A different point is made by **Munsat** in his interesting remark about linguistics. According to Munsat, linguistics is really a bit like mathematics, and therefore the neuron doctrine has no bearing on it. It is true that the status of linguistics is a difficult matter, and there is certainly some plausibility to the claim that languages or grammars are mathematical objects. But it is not obvious from this that linguistics and linguistic explanation is similar to mathematics and mathematical explanation. Moreover, as Alexander George (1989) makes clear, even a Platonist conception of linguistics requires a psychologistic account of psycholinguistics. Our point will be the same if it is restricted to psycholinguistics.

R2. Is the trivial neuron doctrine really trivial?

Many commentators objected to our use of the word "trivial." We were surprised by this. First, as we noted, we did not intend the word trivial to mean uninteresting, false, logically flawed, or logically vacuous. We meant to suggest only that the TND does not make strong predictions about the future course of science and is an uncontroversial view about the practice of science. Nevertheless the word caused alarm in some readers that we might have avoided with a less contentious choice.

Words aside, some of the objections made to our argument focused on the substantive question of whether the TND is, or should be, widely accepted. If it is not, then it is not trivial in our sense. These objections fall into two groups: those who claim that cognitive scientists do not hold the TND and those who claim that philosophers do not hold it. We consider each in turn.

R2.1. The autonomy of cognition. It is pointed out by **Sutton, Hardcastle, Revonsuo,** and **Daniel** that the TND is at odds with a position that they say is, or was, common in cognitive science circles and therefore should not be called trivial. This position, sometimes referred to as the *autonomy of cognition* or the *autonomy of the mental*, says that the level of description of mental function is independent of the neural level. Therefore, no facts about the brain are relevant to understanding mental function, and no doctrine that envisages the theory of the mind as an integration of the psychological and neural sciences could be true. If the TND, which is one such doctrine, is not true, then it cannot be *obviously* true or uncontroversial, which is what our use of "trivial" is meant to convey. Even if the autonomy thesis is not true, if it is or was widely believed, then the TND cannot be trivial in our sense. We addressed this issue only briefly in the target article, and we are pleased to be given the opportunity to answer it here.

The doctrine of the autonomy of the mental can be interpreted in two quite distinct ways corresponding to the two senses of "autonomy" usefully distinguished by **Stone & Davies.** "C-autonomy" is the relation of independence between theories and "F-autonomy" is the relation of irreducibility between theories. The doctrine of the autonomy of the mental can therefore be interpreted either as the strong thesis that the theory of the mind is C-autonomous with respect to (i.e., independent of) the theory of the brain, or as the weaker thesis that the theory of the mind is F-autonomous with respect to (i.e., irreducible to) the theory of the brain. There are thus two distinct claims that **Daniel, Hardcastle, Revonsuo,** and **Sutton** could have in mind – the first that the TND is inconsistent with the C-autonomy of psychology with respect to neurobiology, and the second that it is inconsistent with the F-autonomy of psychology with respect to neurobiology – and our response to their criticism depends on which of these is under discussion.

We take C-autonomy to represent the position we ascribed to certain researchers in the AI community, social constructivists, and others who take theories of the brain to be strictly irrelevant to the understanding of the mind. Notice that this position is very strong indeed because it implies not only that the theory of the mind will be a solely psychological theory, but that neurobiology could not possibly be relevant to the development of that theory. We agree that the TND is inconsistent with this interpretation of the autonomy of the mental, but we do not believe that this interpretation is commonly held. After all, researchers representing mainstream positions in the sciences of the mind have never denied that neurobiology might be rele-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

vant to the study of the mind. Because it is a mistake to delimit *a priori* the domain of investigation that will provide insight into any phenomenon of interest, how *could* they deny that? More importantly, these researchers did not deny that even an autonomous theory of the mind would require an implementation theory to explain how mental function was instantiated in the brain, even if neurobiology played little or no part in the development of that theory. That is to say, they advocated the thesis of F-autonomy, rather than the thesis of C-autonomy. However, the F-autonomy of psychology with respect to neurobiology is *not* inconsistent with the TND because the TND permits a successful theory of the mind to be constituted by a psychological theory together with a neurobiological implementation to which the psychology is irreducible. We disagree, therefore, with **Daniel, Hardcastle, Revonsuo,** and **Sutton** because we claim that the TND is perfectly consistent with one popular interpretation of the doctrine of the autonomy of the mental, and although the TND is inconsistent with a different interpretation of the autonomy thesis, *that* thesis is not at all widespread.

The debate between those who support a neuroscientific approach to cognition and those who support the autonomy of the mental is often characterized as a debate between two radical views, the RND and the thesis that psychology is C-autonomous with respect to neurobiology. But this choice is a false one. There is an obvious conservative position, namely, the TND, that is compatible with a large range of middle positions (see again **Stone & Davies**). Once this middle ground is pointed out, it is clear that many cognitive scientists and neuroscientists implicitly are located at some point within it or that they would choose to be so. And it is equally clear that this is the right position to hold.

R2.2. Issues in the philosophy of mind and language. Both **Daniel** and **Jamieson** write that the TND is inconsistent with positions defended in philosophy of mind and language and therefore cannot be trivial. (For a further discussion of how the argument of the target article interacts with issues in the philosophy of mind, see sect. R5 below). According to Jamieson, the TND is inconsistent with externalism, functionalism, and property dualism. Daniel also cites the case of functionalism.

We agree that the TND is inconsistent with property dualism; the framework of the target article explicitly included a commitment to materialism (as evidenced, for example, in the argument from materialism and naturalism). In any event, both the TND *and* the RND are inconsistent with dualism. Thus, the question of dualism is orthogonal to that distinction and to our concerns.

We do not agree, however, that the TND is inconsistent with functionalism or with externalism. **Jamieson** has analytic functionalism in mind, as **Daniel** does, according to which functionalism is an account of what our ordinary mental concepts are; it is, so to speak, a "theory-theory" approach to our commonsense folk psychology. But a functionalism that provides an analysis of our everyday mental concepts leaves open a number of options when one turns to the scientific theory of what it is that those concepts denote. For example, our everyday concept of belief might be articulated as the state of a person that represents the world as being in a certain way and which, when combined with other states, causes the person to behave in various ways. Such an account of belief largely leaves open the question of what, *as a matter of empirical fact,* the states in question are – linguistic states, computational states, neurobiological states, or some combination. Analytic functionalism thus largely leaves open how a scientific theory of mental states will look, and this means that analytic functionalism is not inconsistent with the TND. Analytic functionalism is a view about concepts; the TND is a view about what best explains, as a matter of fact, what the concepts denote.

The question of analytic functionalism is important also for another reason. In her commentary, **Hardcastle** makes the interesting point that one cannot conjoin biological theories with psychological theories without loss of generality; any such theory would necessarily only apply to creatures with brains. We agree that psychological concepts should apply to creatures that are very different from us in constitution and structure, but we take it that this generality is the goal of a functionalist analysis of mental concepts. The role of scientific psychological theories, in contrast, is to explain how these concepts are instantiated in particular creatures. The TND is about the latter issue, not the former.

We turn now to **Jamieson**'s point about externalism, the claim (roughly) that the individuation of some psychological facts requires reference to the external world. We mentioned this issue in a note in the target article (see also Stoljar & Gold 1998), but we are grateful to Jamieson – and to **Jackson** who makes a similar point in a different context – for giving us the opportunity to explore it further.

Externalism can be understood in at least two different ways. (We are indebted to Frank Jackson for discussion.) In one interpretation of the view, externalism is the thesis that some psychological properties are *relational* and *intrinsic: relational* because in order to describe the properties, one has to refer to items external to the subject that possesses them; *intrinsic* because if a subject has a particular psychological property, a duplicate of the subject will also have the property. An analogy is water-solubility. Being water-soluble is a relational property of a sugar cube because in order to describe the property, one has to refer to items external to the sugar cube. However, water-solubility is intrinsic because (in usual cases) duplicate sugar cubes will both be water-soluble if either is. A different interpretation of externalism takes it to be the thesis that psychological properties are *relational* but *extrinsic: relational* for the reason already mentioned; *extrinsic* because if a subject has a property, it does *not* follow that a duplicate subject will also have the property. An analogy is the property being two feet from a burning barn. If a sugar cube is two feet from a burning barn, it does not follow that a duplicate sugar cube will also be two feet from a burning barn.

Jamieson's claim is that the TND is inconsistent with externalism because neural phenomena are internal; the identity of mental phenomena and neural phenomena to which the TND is committed entails that mental phenomena must be internal as well. Jamieson rightly claims, however, that it is a common view that some mental phenomena have to be individuated by reference to the external world and cannot therefore be internal.

What does it mean to say that neurobiological properties are internal? Usually, this claim is understood as the claim that neurobiological duplicates will have all and only the same neurobiological properties. But that is quite consistent with externalism in the first sense. As long as neurobiological properties can be relational properties, they can provide the connections to the external world required by

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

externalism about mental phenomena. The neurobiological properties will nonetheless remain internal in the sense that they will be preserved across duplication. Moreover, it seems quite clear that some neurobiological properties *are*, in fact, relational in the sense required. (This point is also made by **Perring.**) The concept of the receptive field in visual neurophysiology provides an illustration. Receptive fields are defined by reference to an area of visual space in which a stimulus elicits a response from a neuron and by reference to the particular stimulus that elicits that response. Neurons in V1, for example, respond to moving bars. The classification of V1 cells, therefore, makes essential reference to the external world (i.e., to a moving bar). Externalism in the intrinsic sense, therefore, is not at odds with the TND.

What if one interprets externalism in the extrinsic sense? Admittedly, the claim that the TND is inconsistent with this version of externalism is more plausible than the parallel claim about the first version. If neurobiological properties are internal in the sense of being preserved across duplication but mental properties can *change* across duplication, then mental properties cannot be identical to neural properties in the sense presupposed by the TND. Having distinguished these two senses of externalism, however, it is not clear to what extent this latter view is widely held. (**Jackson**, for example, holds the first version, not the second.) And if it is not widely held, then the TND can still be said to be trivial in our sense. In any case, the *general* claim that externalism is incompatible with the TND is false because of its compatibility with the intrinsic version of externalism.

More importantly, to the extent that the extrinsic version of externalism is plausible in the psychological case, we must ask whether neurobiological properties *are* indeed internal. If both neurobiological and psychological properties can be relational, as we have claimed, why is it plausible to treat only the *latter* as extrinsic? Certainly, an argument to the effect that neurobiological properties are not extrinsic is required before a possible incompatibility with the TND can be said to have been shown.

R3. Is there a distinction between the TND and the RND at all?

We turn next to a different sort of criticism regarding the interpretation of the neuron doctrine, namely, that the distinction we draw between the TND and the RND is not well-founded. This objection is formulated by the commentators based on a number of different premises.

R3.1. Reduction and implementation. According to **Sutton,** we have distinguished too rigidly between reduction and implementation, thereby setting the bar too high for real science. As a result, the distinction between an account of a mental phenomenon that is purely neurobiological and one that is interdisciplinary is not as clear-cut as we maintain. With a less rigid TND–RND distinction and a more nuanced conception of reduction, it is possible to show that Kandel's theory *is* a reduction of classical conditioning.

In the target article, we took for granted that there was an intuitive distinction between reduction and implementation, and although we referred to the classical conception of reduction (Kim 1993; Nagel 1961), we made no specific claims about the proper way to characterize the distinction. No doubt there is a lot of philosophy of science to be done here. However, our argument does not depend on any precise conception of reduction and implementation. The intuitive way to express our view about reduction is that the form of a successful theory depends on where the best explanation of the phenomena is to be found (sect. 5.3.6, target article). Thus, leaving aside the details of reduction and implementation, we claim that the RND is true just in case the best explanations of the phenomena of the mind are to be found in neurobiology.

We agree with **Sutton** that the nature of reduction is relevant to the understanding of the theory of the mind, and we suspect that he is right that mixed positions of the sort he describes better capture the likely future of the sciences of the mind. However, the plausibility of the TND does not by itself show that there will never be an extension of neurobiology that will explain the mind. Because the RND could be true, the distinction between the RND and the TND is a real one.

R3.2. Metaphysical considerations and explanation. The distinction between the TND and the RND, **O'Meara** argues, depends on a distinction between a Humean and a physical conception of causation. According to his view, the TND presupposes a Davidson-style psychophysical supervenience thesis that depends on a Humean account of causation. The RND, in contrast, presupposes a physical account of causation, according to which psychological states and events could not be causal in the Humean sense. Thus, for O'Meara, the TND and the RND are in fact the same claim interpreted according to two different accounts of causation.

O'Meara's distinction does not, however, affect the point we want to make. It is commonplace in discussions of causation to make a distinction between causation proper and causal explanation, where "causation" refers to the metaphysical phenomenon itself and "causal explanation" refers to the process of appealing to that phenomenon in the context of a scientific theory or model. O'Meara's distinction between "physical" and "Humean" theories is a distinction at the metaphysical level and not at the level of scientific explanation. Our argument concerns the level of scientific explanation because it addresses the kind of theory that will offer explanations of mental phenomena. Indeed, these are not necessarily *causal* explanations, because it is far from obvious that all psychological explanations are causal. Whether one thinks that causation is a physical relation or not, one will still have to make sense of the idea that explanations operate at both the psychological and neurobiological levels. Our interest is in the relation between the explanations at these levels.

It is worth pointing out also that **O'Meara**'s suggestion that psychology is operating with a Humean account of causation while neurobiology is operating with a physical account, is not altogether plausible. A claim like "the property *P* of the stimulus caused the perceiver to be in perceptual state *S*" might well be heard in a psychology lab (and similar claims might well be heard in a neurobiology lab). However, the claim itself leaves the issue of the background metaphysics of causation quite open. The difference between the psychological and neurobiological causal stories does not seem to be a difference in causation but a difference between psychology and neurobiology.

In contrast, **Fahey & Zenzen** argue that there *is* a clear TND–RND distinction but that our way of attempting to

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

articulate it is inadequate. We distinguish between explanation by cognitive neuroscience as opposed to explanation by neurobiology alone. But Fahey & Zenzen claim that all explanation is relative to some purpose or other, and the mere availability of an explanation cannot be used as the arbiter of whether reduction has been achieved in some domain or other. The only way to defend reductionism is to be committed to a metaphysical principle according to which the properties of higher-level phenomena are derivable from the properties of lower-level phenomena. To claim that the RND is false, therefore, one must be an emergentist; that is, one must believe that the properties of higher-level entities are *not* so derivable.

Is a metaphysical commitment as strong as emergentism needed to support the TND–RND distinction? We do not think so. Imagine that the science of the mind one day has adequate explanations for all mental phenomena. It seems that one can coherently ask the question, "Are those explanations purely neurobiological or not?" in the absence of any belief whatsoever about emergentism. It may well be that the course of science is in fact affected by whether or not emergentism is true, but that is a quite distinct claim. Because the question about explanations of mental phenomena seems perfectly coherent, it seems the TND–RND is well-founded. It may be true that explanations are purpose-relative, but that does not make it impossible to ask what concepts are used in an explanation.

R3.3. Description and explanation. A different strategy is adopted by **Lamm,** who distinguishes a radical and a trivial doctrine by appealing to "genuine" explanation as opposed to mere description. The genuinely radical doctrine is committed to the explanation of a cognitive phenomena completely in neural terms, whereas the trivial doctrine is only committed to the existence of descriptions of those phenomena. Lamm argues that psychological theories are really only descriptions and thus only support the trivial doctrine. Lamm does not think that this is a criticism of psychology, however, because all sciences must go through a descriptive stage.

Lamm is right in distinguishing different stages in the development of scientific theory, although we hope he is wrong that the best we can expect is a descriptive science of the mind. (In the quotation from the target article that Lamm provides, we *did* only mention description. In surrounding sentences we also mentioned explanation, but the quoted sentence is certainly misleading in that it overemphasizes the issue of description.) Where Lamm sees the explanatory theory of the mind as a largely neurobiological one with psychological concepts thrown in, we believe that an explanatory theory *might* be mostly psychological. We do not need to take a stance on how likely this is, but it is important to keep such a possibility among the serious options for the explanation of the mind.

R3.4. Relations among the sciences. According to **Bullinaria,** we have elevated a pragmatic matter of developing a scientific explanation into a call for more work in a new subject, the philosophy of neuroscience. His claim is that a commitment to reductionism in science only expresses the aim of choosing an appropriate theoretical level that will provide a perspicuous explanation of the phenomenon in question, and thus that the TND–RND distinction is a practical rather than a substantive one.

It is not clear to us, however, why **Bullinaria** takes his view of the pragmatics of explanation to be at odds with our account of the neuron doctrine. We did not explore how particular scientific theories are chosen by scientists to explain particular phenomena. Once such explanations are produced, however, one can ask which theories were used to produce them. Was quantum mechanics used to explain gene replication? If not, then the explanation of gene replication does not come from that level of theory. Is it relevant that genes are made up of fundamental physical particles? Not in the least, unless one can generate a better explanation of gene replication in quantum mechanical terms. (See also our response to **Byrne & Hilbert.**)

Now consider the theory of the mind. Whatever determines the choice of theories that are used to produce explanations of mental phenomena, it is quite coherent to ask, *Were theories other than neurobiological ones required?* If the answer is yes, then the RND is false; if it is no, then the RND is true. The debate over which theories to choose for which purposes of explanation is orthogonal to our position, which is that there are two *real scientific futures* that can be envisaged in the sciences of the mind. In one, various sciences are required to produce good explanations of mental phenomena. In another, neurobiology alone is enough to produce those explanations. One can therefore hold **Bullinaria**'s views about explanation without denying anything we say.

Horwitz also suggests that we have created a dichotomy where there is none. He argues that neuroscience makes use both of neural and higher-level psychological concepts in order to understand cognitive phenomena, and he cites a number of studies from "modern" neuroscience to illustrate the approach. (No doubt Eric Kandel would disagree that his work reflects an outmoded style of neuroscience!).

Horwitz's commentary is clearly meant to be an objection to the argument of the target article, but we suspect that our use of the word "trivial" has distracted him from the substance of our position. So far as we can see, Horwitz's description of neuroscience is precisely one plausible and attractive instantiation of the TND – a truly integrative and interdisciplinary cognitive neuroscience. As we say in the target article and have reiterated above, we did not mean in the least to demean cognitive neuroscience by our use of the word "trivial"; quite the contrary. Our use of "trivial" was meant to imply the obvious truth of the neuron doctrine and the fact that it is and should be widely accepted in the neuroscientific community. We therefore take Horwitz's commentary to support our position rather than to reveal its flaws.

R3.5. Is the confusion caused by a different distinction? We conclude this section by discussing a related objection in the interesting commentary of **Stone & Davies.** They do not deny the distinction between the TND and the RND but argue that it is implausible that these are the positions that are conflated by scientists and philosophers. Instead, they distinguish a weak and a moderate neuron doctrine. According to the former, neurobiology will explain some aspects of the mind and psychology will explain others; according to the latter, neurobiology will be integrated with psychology to explain the mind. The weak doctrine takes the science of the mind to be multidisciplinary, whereas the moderate doctrine takes it to be interdisciplinary. Neither doctrine is equivalent to the TND; Stone & Davies cor-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

rectly claim that the TND is strictly compatible with the idea that neurobiology will play *no* role in the future theory of the mind (see sect. 2.1 above). That is, the TND allows *any* combination of psychological and neurobiological concepts including the "vacuous" cases in which neurobiology either plays no role, or plays the only role. Stone & Davies argue that it is implausible to suppose, as we do, that scientists and philosophers mistake a doctrine as weak as the TND with the RND and much more likely that they conflate the RND with the moderate doctrine.

We agree that the moderate and, indeed, even the weak doctrines are quite plausible views about the future of the sciences of the mind, and it is quite possible that many scientists and philosophers do mean to endorse the moderate view when they endorse the RND. But it is precisely the specificity of this view that makes us doubt the suggestion that it is common to conflate it with the RND. Anyone who had formulated the moderate position would have quite clear ideas about the possible roles of neurobiology and psychology in some future theory of the mind. Our suggestion, is that a general commitment to the identity of mind and brain and a confidence in the ability of science to explain the mind could form the basis for an illicit inference to the RND. In other words it is easier to leap to the RND from the obvious plausibility of the TND than from the specific prediction of the moderate doctrine. To be sure, this is a speculation about the psychology of people who are interested in the neuron doctrine but one that seems to us quite plausible.

R4. Were we sufficiently critical of the RND?

A number of commentators suggest that we have not gone far enough in our criticism of the RND.

Before we address the specific arguments of the commentators, we should say more about our overall view of the RND. In the target article, we claim that a commitment to naturalism is at odds with confident predictions about science. It is our commitment to naturalism that leads us to be agnostic about the future of the science of the mind, and this agnosticism extends to the RND. Whatever the current state of the science of the mind, it seems perverse to deny the *possibility* of a scientific development, however radical the possibility seems, that *will* produce concepts that would allow neurobiology to explain mental phenomena. Our conclusion is not that the RND is false but that there is currently no positive case to be made for it.

The commentary of **Uttal** is a case in point. We cannot possibly do justice to the many interesting suggestions he makes purporting to show the falsity of the RND. One of his claims is that the apparent randomness of the brain shows that it is impossible to work up from neurobiology to mind or backwards from mind to neurobiology. Consider, however, the neurobiological work of Freeman (1991) [See also Skarda & Freeman: "How Brains Make Chaos to Make Sense of the World" *BBS* 10(2) 1987.] who has shown that nonlinear dynamical modelling reveals stable mathematical states in the apparently random activity of neurons in the olfactory bulb of rabbits. These states are not only stable but alter systematically and predictably with changes in olfactory stimulus. Freeman's work thus reveals a surprising fact, namely, that with the appropriate theory, not only can one make sense of the apparent randomness in brain function but one can represent the neural level in such a way as to make contact with psychology. It seems premature, therefore, to conclude that future developments in the science of the mind might not make progress of the same sort; a revolution, or a number of revolutions, might do away with psychology and vindicate the RND. The probability of such a revolution may be low, but we do not believe it is zero.

R4.1. Functional roles. In order to counter this claim, **Chater** compares psychology and economics and argues that just as no physical description of the money in my pocket will explain its role as currency, so no neurobiological description of brain activity will explain psychological function. In each case, the lower-level description picks out the realizer of the functional state but is inadequate to replace the functional description. (See also sect. R2.2 above.)

Notice first that as part of a social institution, currency has physical properties that are stipulated by the relevant authorities. For this reason, **Chater** is quite right that the physical description of money is of no relevance to the functional role played by a particular currency. Because in principle paisley socks could play the role of dollar bills as easily as certain kinds of paper and ink, there is nothing interesting to be learned about the role of currency from the examination of its realizers. But it is by no means obvious that the same can be said for the relation between mind and brain; *not just any* potential realizer will be adequate to do the job defined by the functional description, nor is any functional description compatible with particular realizers. As **Jackson** notes, one's view about the individuation of beliefs can be affected by discoveries about neural function. A comment we made about Kandel in the target article makes the same point. According to Kandel, neurobiology may have discovered that simple and associative learning are not as different as psychology has supposed. Perhaps that insight could have been achieved by psychology alone; perhaps not. The fact that neurobiology can contribute to the functional-level description, however, shows that the possibilities for what Patricia Churchland (1986) calls the "coevolution" of neurobiology and psychology are real. And if neurobiology can make *some* contribution to the functional-level description of mental life, it is hard to see how *a priori* limits on that contribution could be established. In this regard, it is worth reiterating **Lau**'s point that Marr himself acknowledged that there may be cases in which insight into the function of some psychological system may *demand* a nuts-and-bolts explanation rather than a purely psychological description. If there are such cases, then they are evidence that a successful theory of the mind will include a significant conceptual neurobiological component.

Smart, to whose seminal views about the mind-body problem our own perspective is so much indebted, makes a point similar to Chater's in a different way. Smart argues that psychology is to neurobiology as block diagrams are to wiring diagrams. Given the indispensability of broadly functional description to psychology as well as to electronics, he concludes that there is no prospect of neurobiology replacing psychology.

We also think that functional descriptions are indispensable and that most of those descriptions currently come from psychology. Where we differ from **Smart** is in our belief that neurobiology may eventually be able to produce its own adequate functional descriptions of psychological phe-

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

nomena. Should that day come, we could agree with Smart that functional descriptions are an integral part of the science of the mind and still maintain that the RND is true.

Tirassa defends the surprising claim that the RND follows from the computational theory of the mind that is currently the accepted framework of psychology. According to his view, as against those that defend the autonomy of the mental, it is a slippery slope from the distinction between functional role and implementation to a rejection of the importance of the functional description. If psychology is no more than an abstract description of the brain, why not simply study the brain? Tirassa defends a conception of minds as ontologically primitive, from which it follows that psychology is ineliminable.

Tirassa may of course be right in thinking that the current division of labor between computational description and neurobiological implementation may one day change, but it is unclear to us how the ontological primitiveness of minds leads to the belief that psychology is ineliminable. The connections between general metaphysical views and the development of scientific theories are much less obvious than is often thought. As a result, even if we did take minds to be primitive parts of our ontology, nothing substantive follows for the sciences of the mind. Perhaps the best description of a mind-as-primitive is a neurobiological description. There is thus no difficulty in holding that the RND may eventually be vindicated, whatever the background metaphysics.

R4.2. Unification. An attempt is made by **Franceschetti** to show that we have been hoist by our own petard by arguing that considerations of unification not only do not provide evidence for the RND, they provide evidence against it. He argues that, in practice, theories tend to co-evolve, and that even in cases of genuine unification (e.g., by Maxwell's equations), theories apparently made otiose (e.g., optics) continue to remain useful. Thus even if there were a unification of some sort in the sciences of the mind, psychology is unlikely to disappear and the RND is unlikely to be true. (This argument, if sound, would lend support to the position of **Byrne & Hilbert** with respect to unification.)

Our response to **Bullinaria** is relevant here. The target article did not discuss the factors that determine where a successful scientific explanation will be found. Our position is only that once an explanation is available, there is a determinate answer to the question of which science or sciences produced it. If optics is still required to explain particular phenomena, then Maxwell's equations did not unify optics, if by unification one means dissolution. Similarly, should a unified theory of the mind be produced, it will either include psychology as a necessary part, or it will not. If it turns out that psychology is unnecessary (even if it remains a convenient shorthand), then the RND will have been vindicated.

R4.3. On the use and abuse of neuroscience. We are grateful for **Perring**'s commentary, in which he supports some elements of our position by appealing to the case of psychiatry. He shows that the TND and the RND are regularly conflated, with potentially serious consequences. An advocate of the neuron doctrine who fails to make the TND–RND distinction will infer from a "mental illness is an illness of the brain" that "mental illness ought to be treated by neurobiological interventions." But, as Perring notes, this does not follow at all, as talk therapy, for example, may continue to be the most effective treatment for a pathology of the brain. Given the enormous emotional, social, and financial costs of mental illness, if Perring is right, then the TND–RND distinction cannot be of purely intellectual interest. The commentary of **Brothers** also makes reference to the social effects of the RND. Brothers argues that the RND has received unreasonable support in part because of the social value of trumpeting the successes of neuroscience in particular at a time when global theories of the mind (e.g., Freud's) are found sorely lacking. Once the social motivation behind the RND is noticed, however, its appeal is significantly reduced.

Is the RND a remedy for a *fin de siècle* malaise in the scientific community? Perhaps; the question is difficult and we don't have the expertise to answer it. What is important to stress, however, is that whatever the uses and abuses of the RND may be, one must keep its possible truth separate from them. However misguided some advocates of the RND might be, they may be advocating a view that is eventually vindicated. This is a theme familiar in other areas of science as well. For example, few theories are as badly misrepresented and abused as quantum mechanics, as a cursory look in the "metaphysics" section of many local bookstores will reveal. But the use of quantum mechanics to support everything from panpsychism to Eastern medicine does nothing to negate the theory. Likewise, one ought to remain open to the RND whatever its current social abuses.

R5. How does our argument bear on related issues in the ontology of mind?

As noted in **Perring**'s commentary, our aim in the target article is to deal with a number of issues in the philosophy of science and their implications for neuroscience. However, some commentators raised related questions about the ontology of the mind that we did not deal with explicitly. We conclude with a discussion of some of these issues.

Jordan suggests that the naturalism we are committed to is at odds with a commitment to materialism. He suggests, in effect, that an open-minded investigation into the nature of the mind ought to refrain from commitments unsupported by science, and materialism is one such commitment. Opting for neutral monism – according to which the mind and the brain are aspects of the same neutral stuff – is, therefore, preferable.

Our commitment to materialism for the purposes of the target article, however, amounts to no more than the claim that the mind is identical to the brain; indeed, even this is only a working assumption. Our intention is to suggest that ontological monism may coexist with scientific pluralism, and we could have adopted a neutral monism to make the same point. The dispute between neutral monism and materialism is separate from the main claims we defend.

Mogi acknowledges that the RND is radical but offers an interesting suggestion as to why a theory of the mind should invoke explanations at the individual neural level. He then raises the problem of qualia. As we understand it, his view is much like Barlow's (1987), which we referred to in the target article. According to Barlow, biological neuroscience is promising as an explanation of mental phenomena, at least until we reach the problem of conscious awareness. Like **O'Meara,** Mogi advocates a supervenience account as

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

both potentially adequate to the phenomena and capable of being integrated into a respectable science of the mind.

We are rather doubtful about the potential usefulness of supervenience in this context for two reasons. First, we are impressed by the work of Kim (1993), who argues that the notion of a wholly independent psychological domain that nonetheless supervenes on a distinct neurobiological domain may be untenable. What is more relevant to the argument of the target article, however, is that the supervenience of the psychological on the neurobiological or physical is usually taken to be a metaphysical thesis, not an explanatory one. Our interest is in the explanatory question, not the metaphysical one.

Partly because of this, we do not say anything in the target article about qualia or consciousness more generally, nor do we address in any direct way the mind–body problem or the problem of other minds, both raised by **Gunderson.** It is important, however, to have something to say about how these issues are related to the target article. With respect to the mind–body problem, the crucial question is not whether neuroscience or cognitive psychology will be able to explain mental phenomena such as qualia. It is whether a science of *any* sort will be able to do so (see Jackson 1982). So the mind–body problem raises questions about the scope and limits of science in general, and not simply about neuroscience and cognitive psychology. Of course, these are the sciences that we look to for a solution to the mind–body problem, so they bear the burden of our search.

With respect to qualia and consciousness in general, one of the obstacles to an adequate theory is the absence of an adequate functional characterization of consciousness (Block 1995). Of course, it is plausible that some of the phenomena connoted by the concept of consciousness, for example, what Block calls *access-consciousness*, do have functional characterizations. But, as Block points out, it does not follow from this that we have a functional characterization of qualia or phenomenal consciousness. As a result, we are inclined to be pessimistic about the likely success of the current search for the neural substrate of consciousness (e.g., synchronous 40-Hz oscillations). Until we have a better idea of what we are looking for, it is going to be difficult to find it.

As for the problem of other minds, this seems to us a normative question in epistemology about which the descriptive questions in philosophy of science have nothing in particular to say. **Gunderson** notes that we acknowledge the possibility that not all problems about the mind will be explained by neuroscience. The problem of other minds strikes us as one such case.

Finally, we come to the commentary of **Jackson.** We have already noted his point about externalism, so here we will concentrate on his suggestion concerning the relevance of neuroscience to the ontology of the mind, with which we are much impressed and quite in agreement. Jackson argues that because psychological states, such as belief, are information-bearing, a successful neuroscience will tell us how they are achieved by the brain. In so doing, neuroscience will have something to say about how these states are *individuated* and will thereby make a contribution to mental ontology. To take his example, results in neurobiology might make a difference to whether or not we approach individual beliefs as discrete countable states of the subject or rather as abstractions from a single overall belief state.

Jackson's suggestion is important for two reasons. First, it shows in a particularly convincing way that where questions about mental life depend on the fundamental building blocks of mental representation, we have a strong reason to expect the coevolution of psychology and neuroscience described by Patricia Churchland (1986). Just as the discovery of the double helix had important consequences for larger questions in genetics, the discovery of the properties of information-bearing states in the brain will affect how we count, and conceive of, psychological states. Second, his suggestion highlights the fact that even if discoveries in neuroscience do not undermine the psychological conception of the mind, they may nevertheless alter it in significant ways. How neurons produce mental states is a problem of implementation but – to borrow Gary Marcus's turn of phrase – not *mere* implementation. How the brain does the work of the mind is one of the deep questions of science even if the science of the mind turns out to include psychology.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Amaral, D. G., Insausti, R. & Cowan, W. M. (1984) The commissural connections of the monkey hippocampal formation. *Journal of Comparative Neurology* 224:307–36. [DWZ]

Arieli, A., Serkin, A. & Grinvald, A. (1996) Dynamics of ongoing activity: Explanation of the large variability in evoked cortical response. *Science* 273:1868–71. [SH]

Bailey, C. H. & Kandel, E. R. (1995) Molecular and structural mechanisms underlying long-term memory. In: *The cognitive neurosciences,* ed. M. Gazzaniga. MIT Press. [aIG]

Barinaga, M. (1996) Neurons put the uncertainty into reaction times. *Science* 274:344. [SH]

Barkow, J. H., Cosmides, L. & Tooby, J. (1992) *The adapted mind: Evolutionary psychology and the generation of culture.* Oxford University Press. [rIG]

Barlow, H. B. (1972) Single units and sensation: A neuron doctrine for perceptual psychology? *Perception* 1:371–94. [aIG, KM, JMZ]

(1987) The biological role of consciousness. In: *Mindwaves,* ed. C. Blakemore & S. Greenfield. Blackwell. [arIG]

(1995) The neuron doctrine in perception. In: *The cognitive neurosciences,* ed. M. Gazzaniga. MIT Press. [aIG]

Bear, M. F. & Abraham, W. C. (1996) Long-term depression in hippocampus. *Annual Review of Neuroscience* 19:437–62. [RMV]

Bechtel, W. (1994) Levels of description and explanation in cognitive science. *Minds and Machines* 4:1–25. [AR]

Bechtel, W. & Richardson, R. C. (1993) *Discovering complexity: Decomposition and localization as strategies in scientific research.* Princeton University Press. [AR]

Beck, F. & Eccles, J. C. (1992) Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Sciences USA* 89(23):11357–61. [SH]

Bickle, J. (1998) *Psychoneural reduction: The new wave.* MIT Press. [JS]

Blackburn, S. (1991) Losing your mind: Physics, identity and folk burglar prevention. In: *The future of folk psychology,* ed. L. Greenwood. Cambridge University Press. [SGD]

Bliss, T. V. P. & Lømo, T. (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* 232:357–74. [NC, aIG]

Block, N. (1990) The computational model of the mind. In: *Language: An invitation to cognitive science,* ed. D. N. Osherson & H. Laznik. MIT Press. [aIG]

(1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18(2):227–87. [rIG]

Bogen, J. E. (1992) The callosal syndromes. In: *Clinical neuropsychology,* ed. K. M. Heilman & E. Valenstein. Oxford University Press. [DWZ]

Boring, E. G. (1950) *A history of experimental psychology,* 2nd edition. Appleton-Century-Crofts. [aIG]

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

Brothers, L. (1997) *Friday's footprint: How society shapes the human mind.* Oxford University Press. [LB]
Bullinaria, J. A. (1986) Chiral fermions in Kaluza-Klein theory. *Nuclear Physics* B272:266–80. [JAB]
Buzsáki, G. (1989) Two-stage model of memory trace formation: A role for "noisy" brain states. *Neuroscience* 31(3):551–70. [aIG]
Castellucci, V. & Kandel, E. R. (1976) Presynaptic facilitation as a mechanism for behavioral sensitization in *Aplysia. Science* 194:1176–78. [aIG]
Chalmers, D. (1996) *The conscious mind.* Oxford University Press. [aIG, KM, JMZ]
Churchland, P. M. (1980) Critical notice: Joseph Margolis, *Person and minds: The prospects for nonreductive materialism. Dialogue* 19:461–69. [aIG]
(1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78:67–90. [aIG]
(1989a) *A neurocomputational perspective: The nature of mind and the structure of science.* MIT Press. [aIG]
(1989b) On the nature of theories: A neurocomputational perspective. In: *A neurocomputational perspective: The nature of mind and the structure of science.* MIT Press. [aIG]
(1989c) Some reductive strategies in cognitive neurobiology. In: *A neurocomputational perspective.* MIT Press. [aIG]
(1990) Cognitive activity in artificial neural networks. In: *Thinking: An invitation to cognitive science,* ed. D. N. Osherson & H. Lasnik. MIT Press. [aIG]
(1995) *The engine of reason, the seat of the soul: A philosophical journey into the brain.* Cambridge University Press. [SGD, aIG, SH]
(1996) Flanagan on moral knowledge. In: *The Churchlands and their critics,* ed. R. McCauley. Blackwell. [JS]
Churchland, P. M. & Churchland, P. S. (1990) Intertheoretic reduction: A neuroscientist's field guide. *The Neurosciences* 2:249–56. [TS]
(1994) Intertheoretic reduction: A neuroscientist's field guide. In: *The mind-body problem,* ed. R. Warner & T. Szubka. Blackwell. [AB, aIG, DJ]
(1996a) Replies from the Churchlands. In: *The Churchlands and their critics,* ed. R. N. McCauley. Blackwell. [aIG, TS]
(1996b) The future of psychology, folk and scientific. In: *The Churchlands and their critics,* ed. R. McCauley. MIT Press. [JS]
(1998) Recent work on consciousness: Philosophical, theoretical, and empirical. In: *On the contrary: Critical essays, 1987–1997.* MIT Press. [JS]
Churchland, P. S. (1986) *Neurophilosophy: Toward a unified science of the mind/brain.* MIT Press. [SGD, arIG, SH, VGH, TS]
(1996) Feeling reasons. In: *The neurobiology of decision-making,* ed. A. R. Damasio, H. Damasio & Y. Christen. Springer-Verlag. [JS]
(1997) Can neurobiology teach us anything about consciousness? In: *The nature of consciousness,* ed. N. Block, O. Flanagan & G. Güzeldere. MIT Press. [AB, rIG]
Churchland, P. S. & Ramachandran, V. S. (1993) Filling in: Why Dennett is wrong. In: *Dennett and his critics,* ed. B. Dahlbom. Blackwell. [JS]
Churchland, P. S. & Sejnowski, T. J. (1992) *The computational brain.* MIT Press. [aIG, SH, CL, JS]
Clarke, E. & Jacyna, L. S. (1987) *Nineteenth century origins of neuroscientific concepts.* University of California Press. [aIG]
Collins, R. (1998) *The sociology of philosophies: A global theory of intellectual change.* Harvard University Press. [LB]
Coltheart, M. & Langdon, R. (1998) Autism, modularity and levels of explanation in cognitive science. *Mind and Language* 13:138–52. [TS]
Crick, F. (1994) *The astonishing hypothesis.* Scribners. [aIG, CL, JMZ]
Cummins, R. (1983) *The nature of psychological explanation.* MIT Press. [aIG]
Davidson, D. (1970) Mental events. In: *Experience and theory,* ed. L. Foster & J. Swanson. Humanities Press. Reprinted in: *Essays on action and events.* Oxford University Press (1980). [KM]
(1980) *Essays on actions and events.* Clarendon Press. [JTO]
Del Re, G. (1998) Ontological status of molecular structure. *HYLE, An International Journal for the Philosophy of Chemistry* 4:81–103. [AR]
Denk, W., Yuste, R., Svoboda, K. & Tank, D. W. (1996) Imaging calcium dynamics in dendritic spines. *Current Opinion in Neurobiology* 6:372–78. [RMV]
DeRenzi, E., Faglioni, P. & Spinnler, H. (1968) The performance of patients with unilateral brain damage on face recognition tasks. *Cortex* 5:274–84. [DWZ]
DeZazzo, J. & Tully, T. (1995) Dissection of memory formation: From behavioral pharmacology to molecular genetics. *Trends in Neuroscience* 18:212–18. [aIG]
De Zeeuw, C. I., Hertzberg, E. L. & Mognaini, E. (1995) The dendritic lamellar body: A new neuronal organelle putatively associated with dendrodendritic gap junctions. *Journal of Neuroscience* 15(2):1587–604. [SH]
Dominey, P. F. & Arbib, M. A. (1992) A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex* 2:153–75. [BH]
Dowe, P. (1992) Wesley Salmon's process theory of causality and the conserved quantity theory. *Philosophy of Science* 59:195–216. [JTO]
(1995) Causality and conserved quantities: A reply to Salmon. *Philosophy of Science* 62:321–33. [JTO]
Dretske, F. (1995) *Naturalizing the mind.* MIT Press. [AR]
Duff, M. J. (1998) A layman's guide to M theory. Paper presented at the Abdus Salam Memorial Meeting, ICTP, Trieste, November 1997. [http://xxx.soton.ac.uk/abs/hep-th/9805177]. [JAB]
Dupré, J. (1993) *The disorder of things: Metaphysical foundations of the disunity of science.* Harvard University Press. [aIG]
Eccles, J. C. (1992) Evolution of consciousness. *Proceedings of the National Academy of Science USA* 89(16):7320–24. [SH]
Edelman, G. M. (1989) *The remembered present.* Basic Books. [aIG]
Enc, B. (1983) In defense of the identity theory. *Journal of Philosophy* 80:279–98. [JS]
Fodor, J. (1981) Special sciences. In: *Representations.* MIT Press. [aIG]
(1989) Making mind matter more. *Philosophical Topics* 17:59–80. [TS]
(1994) Special sciences (or: The disunity of science as a working hypothesis). Reprinted in: *Readings in the philosophy of social science,* ed. M. Martin & L. C. McIntyre. MIT Press. [JTO]
(1998) *In critical condition.* MIT Press. [aIG, TS]
Fodor, J. & Pylyshyn, Z. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3–71. [aIG]
Freeman, W. J. (1991) The physiology of perception. *Scientific American* 264(2):78–85. [rIG]
Frost, W. N., Castellucci, V. F., Hawkins, R. D. & Kandel, E. R. (1985) Monosynaptic connections from the sensory neurons of the gill- and siphon-withdrawal reflex in *Aplysia* participate in the storage of long-term memory for sensitization. *Proceedings of the National Academy of Science USA* 82:8266–69. [aIG]
Funahashi, S., Chafee, M. V. & Goldman-Rakic, P. S. (1993) Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature* 365:753–56. [BH]
Gazzaniga, M., ed. (1995) *The cognitive neurosciences.* MIT Press. [aIG, CL, AR]
(1997) *Conversations in the cognitive neurosciences.* MIT Press. [AR]
George, A. (1989) How not to be confused about linguistics. In: *Reflections on Chomsky,* ed. A. George. Blackwell. [rIG]
Georgopoulos, A. P., Lurito, J. T., Petrides, M., Schwartz, A. B. & Massey, J. T. (1989) Mental rotation of the neuronal population vector. *Science* 243:234–36. [aIG]
Geschwind, N. & Galaburda, A. M. (1984) *Cerebral dominance: The biological foundations.* Harvard University Press. [DWZ]
Gierer, A. (1983) Relation between neurophysiological and mental states: Possible limits of decodability. *Naturwissenschaften* 70:282–87. [JMZ]
Gleick, J. (1992) *Genius: The life and science of Richard Feynman.* Pantheon Press. [aIG]
Gluck, M. A. & Thompson, R. F. (1987) Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review* 94:176–91. [aIG]
Gray, C. M., König, P., Engel, A. K. & Singer, W. (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338:334–37. [aIG]
Guterman, L. (1998) I react therefore I am. *New Scientist* 21:34–37. [AR]
Hall, S. S. (1998) Our memories, our selves. *New York Times Magazine,* February 15. [aIG]
Hameroff, S. R. (1998a) "Funda-mentality" - Is the conscious mind subtly connected to a basic level of the universe? *Trends in Cognitive Science* 2(4):119–27. [SH]
(1998b) 'More neural than thou' (A reply to Patricia Churchland). In: *Toward a science of consciousness II - The Second Tucson Discussions and Debates,* ed. S. R. Hameroff, A. W. Kaszniak & A. C. Scott. MIT Press. [SH]
(1998c) Quantum computation in brain microtubules? The Penrose-Hameroff 'Orch OR' model of consciousness. *Philosophical Transactions of the Royal Society of London A* 356(1743):1869–96. [SH]
Hameroff, S. R. & Penrose, R. (1996a) Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. In: *Toward a science of consciousness - The First Tucson Discussions and Debates,* ed. S. R. Hameroff, A. W. Kaszniak & A. C. Scott. MIT Press. [SH]
(1996b) Conscious events as orchestrated spacetime selections. *Journal of Consciousness Studies* 3(1):36–53. [SH]
Hardcastle, V. G. (1996) *How to build a theory in cognitive science.* SUNY Press. [VGH]
Hardin, C. L. (1988) *Color for philosophers.* Hackett. [aIG]
Hawkins, R. D., Abrams, W., Carew, T. J. & Kandel, E. R. (1983) A cellular mechanism of classical conditioning in *Aplysia:* Activity-dependent presynaptic facilitation. *Science* 219(4583):400–405. [aIG]
Hawkins, R. D. & Kandel, E. R. (1984) Is there a cell-biological alphabet for simple forms of learning? *Psychological Review* 91:376–91. [aIG, WRU]
Hawkins, R. D., Kandel, E. R. & Siegelbaum, S. A. (1993) Learning to modulate transmitter release: Themes and variations in synaptic plasticity. *Annual Review of Neuroscience* 16:625–65. [aIG]

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

Haxby, J. V., Ungerleider, L. G., Horwitz, B., Rapoport, S. I. & Grady, C. L. (1995) Hemispheric differences in neural systems for face working memory: A PET-rCBF study. *Human Brain Mapping* 3:68–82. [BH]
Hebb, D. O. (1945) *Organization of behaviour: A neuropsychological theory.* Wiley. [aIG]
Heiligenberg, W. (1991) The jamming avoidance response of the electric fish, *Eigenmannia:* Computational rules and their neuronal implementation. *Seminars in the Neurosciences* 3:3–18. [aIG]
Higginbotham, J. (1990) Philosophical issues in the study of language. In: *Language: An invitation to cognitive science,* ed. D. N. Osherson & H. Laznik. MIT Press. [aIG, MT]
Hirokawa, N. (1991) Molecular architecture and dynamics of the neuronal cytoskeleton. In: *The neuronal cytoskeleton,* ed. R. Burgoyne. Wiley-Liss. [SH]
Hobson, J. A. (1990) Sleep and dreaming. *Journal of Neuroscience* 10(2):371–82. [aIG]
Horwitz, B. (1998) Using functional brain imaging to understand human cognition. *Complexity* 3:39–52. [BH]
Hubel, D. (1974) Neurobiology: A science in need of a Copernicus. In: *The heritage of Copernicus: Theories "pleasing to the mind,"* ed. J. Neyman. MIT Press. [aIG]
(1988) *Eye, brain, and vision.* W. H. Freeman. [aIG]
Hurvich, L. (1981) *Color vision.* Sunderland. [aIG]
Jackson, F. (1982) Epiphenomenal qualia. *Philosophical Quarterly* 32:127–36. [arIG]
Kandel, E. R. (1987) Preface: Molecular neurobiology and the proper study of humankind. In: *Molecular neurobiology in neurology and psychiatry,* ed. E. R. Kandel. Raven Press. [JS]
Kandel, E. R. & Schwartz, J. H. (1982) Molecular biology of learning: Modulation of transmitter release. *Science* 218(4571):433–43. [aIG, CL]
Kandel, E. R., Schwartz, J. H. & Jessell, T. M. (1995) *Essentials of neural science,* 3rd edition. Wiley. [aIG]
Kandel, E. R., Siegelbaum, S. A. & Schwartz, J. H. (1991) Synaptic transmission. In: *Principles of neural science,* 3rd edition, ed. E. R. Kandel, J. H. Schwartz & T. M. Jessell. Elsevier. [SH]
Kim, J. (1992) "Downward causation" in emergent and nonreductive physicalism. In: *Emergence or reduction? Essays on the prospects of nonreductive physicalism,* ed. A. Beckermann, H. Flohr & J. Kim. De Gruyter. [JF]
(1993) *Mind and supervenience.* Cambridge University Press. [arIG]
King-Smith, P. E. (1991) Chromatic and achromatic visual systems. In: *The perception of colour,* ed. P. Gouras. CRC Press. [aIG]
Konishi, M. (1995) Neural mechanisms of auditory image formation. In: *The cognitive neurosciences,* ed. M. Gazzaniga. MIT Press. [aIG]
Kosslyn, S. M. & Andersen, R. A., eds. (1992) *Frontiers in cognitive neuroscience.* MIT Press. [aIG]
Kosslyn, S. M. & Koenig, O. (1995) *Wet mind: The new cognitive neuroscience.* The Free Press. [aIG]
Lewis, D. (1994) Reduction of mind. In: *A companion to the philosophy of mind,* ed. S. Guttenplan. Blackwell. [FJ]
Livingstone, M. & Hubel, D. (1987) Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience* 7(11):3416–68. [aIG]
Maniotis, A. J., Chen, C. S. & Ingber, D. I. (1997) Demonstration of mechanical connections between integrins, cytoskeletal filaments, and nucleoplasm that stabilize nuclear structure. *Proceedings of the National Academy of Science USA* 94:849–54. [SH]
Manning, A. & Dawkins, M. S. (1992) *An introduction to animal behaviour,* 4th edition. Cambridge University Press. [aIG]
Marr, D. (1982) *Vision.* W. H. Freeman. [aIG, CL]
Marr, D. C. (1977) Artificial intelligence – a personal view. *Artificial Intelligence* 9:37–48. [JYFL]
Mauldin, T. (1996) On the unification of physics. *Journal of Philosophy* 93:129–44. [DRF, aIG]
McCauley, R. M. (1996) *The Churchlands and their critics.* Blackwell. [aIG, JYFL]
Mehler, J., Morton, J. & Jusczyk, P. W. (1984) On reducing language to biology. *Cognitive Neuropsychology* 1:83–116. [TS]
Micevych, P. E. & Abelson, L. (1991) Distribution of mRNAs coding for liver and heart gap junction protein in the rat central nervous system. *Journal of Comparative Neurology* 305:96–118. [SH]
Mogi, K. (1999) Response selectivity, neuron doctrine, and Mach's principle in perception. In: *Understanding representation in the cognitive sciences,* ed. A. Riegler & M. Peschl. Plenum Press. [KM]
Nagel, E. (1961) *The structure of science.* Harcourt, Brace and World. [arIG, JTO]
O'Meara, J. T. (1997) Causation and the struggle for a science of culture. *Current Anthropology* 38(3):399–418. [JTO]
(1999a) Causal individualism and the unification of anthropology. In: *Anthropological theory in North America,* ed. E. Cerroni-Long. International Council of Anthropological and Ethnological Sciences. Bergin & Grove (in press). [JTO]
(1999b) Causation and the application of Darwinian thinking to the explanation of human thought and other behavior (submitted). [JTO]
Oppenheim, P. & Putnam, H. (1958) Unity of science as a working hypothesis. In: *Minnesota studies in the philosophy of science.* University of Minnesota Press. [aIG]
Papineau, D. (1993) *Philosophical naturalism.* Blackwell. [SGD]
Penrose, R. (1989) *The emperor's new mind: Concerning computers, minds, and the laws of physics.* Oxford University Press. [DRF]
(1994) *Shadows of the mind: Concerning computers; an approach to the missing science of consciousness.* Oxford University Press. [DRF, JS]
Penrose, R. & Hameroff, S. R. (1995) What gaps? Reply to Grush and Churchland. *Journal of Consciousness Studies* 2(2):99–112. [SH]
Penrose, R., Shimony, A., Cartwright, N. & Hawking, S. (1997) *The large, the small and the human mind: Concerning computers, minds, and the laws of physics.* Cambridge University Press. [DRF]
Perring, C. (1996) Prozac and political activism. *Perspectives: A Mental Health Magazine* 1:4. (Available at http://www.cmhc.com/perspectives/articles/art09961.htm. September 15, 1996). [CP]
(1998) The rise of philosophy of psychiatry. *Philosopher's Magazine* 3:46–47. (Available at http://www.philosophers.co.uk/current/perring.htm). [CP]
Pinker, S. (1991) Rules of language. *Science* 253:530–35. [aIG]
Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–193. [aIG]
Pinsker, H. M., Kupfermann, I., Castellucci, V. & Kandel, E. R. (1970) Habituation and dishabituation of the gill-withdrawal reflex in *Aplysia. Science* 167:1740–42. [aIG]
Posner, M. I. & Raichle, M. E. (1994) *Images of mind.* W. H. Freeman. [aIG, CL]
Pribram, K. H. (1991) *Brain and perception.* Erlbaum. [SH]
Putnam, H. (1975) Philosophy and our mental life. In: *Mind, language and reality: Philosophical papers, vol. 2.* Cambridge University Press. [aIG]
Pylyshyn, Z. W. (1990) Computation and cognition: Issues in the foundations of cognitive science. In: *Foundations of cognitive science,* ed. J. L. Garfield. Paragon. [AR]
Rasmussen, S., Karampurwala, H., Vaidyanath, R., Jensen, K. S. & Hameroff, S. (1990) Computational connectionism within neurons: A model of cytoskeletal automata subserving neural networks. *Physica D* 42:428–49. [SH]
Rausch, R. & Babb, T. L. (1993) Hippocampal neuron loss and memory scores before and after temporal lobe surgery for epilepsy. *Archives of Neurology* 50:812–17. [DWZ]
Reber, A. S. (1985) *Dictionary of psychology.* Penguin Books. [CL]
Rescorla, R. A. (1968) Probability of a shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology* 66:1–5. [aIG, RMV]
(1980) *Pavlovian second-order conditioning: Studies in associative learning.* Erlbaum. [aIG]
(1988) Pavlovian conditioning: It's not what you think. *American Psychologist* 43:151–60. [arIG]
Rescorla, R. A. & Wagner, A. R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In: *Classical conditioning II: Current research and theory,* ed. A. H. Black & W. F. Prokasy. Appleton-Century-Crofts. [aIG]
Restivo, S. (1985) *The social relations of physics, mysticism, and mathematics.* Reidel. [LB]
Revonsuo, A. (1994) In search of the science of consciousness. In: *Consciousness in philosophy and cognitive neuroscience,* ed. A. Revonsuo & M. Kamppinen. Erlbaum. [AR]
Rosene, D. L. & Van Hoesen, G. W. (1987) The hippocampal formation of the primate brain. In: *Cerebral cortex,* ed. E. G. Jones & A. Peters. Plenum Press. [DWZ]
Russell, B. (1970) *An outline of philosophy.* Allen and Unwin. [JSJ]
Salmon, W. C. (1984) *Scientific explanation and the causal structure of the world.* Princeton University Press. [JTO]
(1994) Causality without counterfactuals. *Philosophy of Science* 61:297–312. [JTO]
Santosh, A. H. (in press) On the possibility of universal neural coding of subjective experience. *Consciousness and Cognition* 8(9). [KM]
Sargent, P. (1996) On the use of visualizations in the practice of science. *Philosophy of Science (Proceedings)* 63:S230–38. [AR]
Sarter, M., Berntson, G. G. & Cacioppo, J. T. (1996) Brain imaging and cognitive neuroscience: Toward strong inference in attributing function to structure. *American Psychologist* 5:13–21. [CL]
Sass, K. J., Lenz, T., Westerveld, M., Novelly, R. A., Spencer, D. D. & Kim, J. H. (1991) The neural substrate of memory impairment demonstrated by the

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

intracarotid amobarbital procedure. *Archives of Neurology* 48:48–52. [DWZ]
Sass, K. J., Spencer, D. D., Kim, J. H., Westerveld, M., Novelly, R. A. & Lenz, T. (1990) Verbal memory impairment correlates with hippocampal pyramidal cell density. *Neurology* 40:1694–97. [DWZ]
Schaffner, K. F. (1992) Philosophy of medicine. In: *Introduction to the philosophy of science,* ed. M. H. Salmon, J. Earman, C. Glymour & J. Lennox. Prentice Hall. [JS]
Scheibel, A. M. (1984) A dendritic correlate of human speech. In: *Cerebral dominance: The biological foundations,* ed. N. Geschwind & A. M. Galaburda. Harvard University Press. [DWZ]
Searle, J. R. (1992) *The rediscovery of the mind.* MIT Press. [AR]
Sellars, W. (1963) Science and the manifest image of man. In: *Science, perception, and reality.* Routledge and Kegan Paul. [aIG]
(1971) Science, sense impressions, and sensa: A reply to Cornman. *Review of Metaphysics* 23:391–447. [aIG]
Shallice, T. (1988) *From neuropsychology to mental structure.* Cambridge University Press. [TS]
Shepherd, G. M. (1991) *Foundations of the neuron doctrine.* Oxford University Press. [aIG, JMZ]
(1994) *Neurobiology,* 3rd edition. Oxford University Press. [aIG]
Sherrington, C. S. (1951) *Man on his nature,* 2nd edition. Cambridge University Press. [SH]
Simon, H. A. (1996) *The sciences of the artificial,* 3rd edition. MIT Press. [DRF]
Simpson, G. V., Pflieger, M. E., Foxe, J. J., Ahlfors, S. P., Vaughan, H. G., Hrabe, J., Ilmoniemi, R. J. & Lantos, G. (1995) Dynamic neuroimaging of brain function. *Journal of Clinical Neurophysiology* 12:432–49. [BH]
Smart, J. J. C. (1963) *Philosophy and scientific realism.* Routledge and Kegan Paul. [JJCS]
(1989) *Our place in the universe.* Blackwell. [JJCS]
Snyder, S. H. (1996) *Drugs and the brain.* W. H. Freeman. [aIG, CL]
Sperry, R. W. (1974) Lateral specialization in the surgically separated hemispheres. In: *The neurosciences third study program,* ed. F. O. Schmitt & F. G. Worden. MIT Press. [DWZ]
Spillman, L. & Werner, J. S. (1990) *Visual perception: The neurological foundations.* Academic Press. [JMZ]
Stoljar, D. & Gold, I. (1998) On biological and cognitive neuroscience. *Mind and Language* 13:110–31. [arIG, JS]
Sutton, J. (1998) *Philosophy and memory traces: Descartes to connectionism.* Cambridge University Press. [JS]
(1999) Distributed memory, coupling, and history. In: *Proceedings of the Fourth Australasian Cognitive Science Society,* ed. R. Heath. [JS]
Tagamets, M.-A. & Horwitz, B. (1998) Integrating electrophysiological and anatomical experimental data to create a large-scale model that simulates a delayed match-to- sample human brain imaging study. *Cerebral Cortex* 8:310–20. [BH]
Tononi, G., Sporns, O. & Edelman, G. M. (1992) Reentry and the problem of integrating multiple cortical areas: Simulation of dynamic integration in the visual system. *Cerebral Cortex* 2:310–35. [BH]
Touchette, N. (1996) You must remember this. *New Scientist (Australia)* 151(2037):32–35. [aIG]
Turner, J. (in press) The neurology of emotions: Implications for sociological theories and interpersonal behavior. In: *Social perspectives on emotion, vol. 5: Minds, brains, and society: Towards a neurosociology,* ed. D. Franks. JAI Press. [LB]
Uttal, W. R. (1998) *Towards a new behaviorism: The case against perceptual reductionism.* Erlbaum. [WRU]
Van Gelder, T. (1995) What might cognition be, if not computation? *Journal of Philosophy* 92:345–81. [JS]
Wagner, A. R. & Pfautz, P. L. (1978) A bowed serial-position function in habituation of sequential stimuli. *Animal Learning and Behavior* 6(4):395–400. [aIG]
Waldrop, M. M. (1993) Cognitive neuroscience: A world with a future. *Science* 261:1805–807. [CL]
Walters, E. T. & Byrne, J. H. (1983) Associative conditioning of single sensory neurons suggests a cellular mechanism for learning. *Science* 219(4583):405–408. [aIG]
Walters, E. T., Carew, T. J. & Kandel, E. R. (1981) Associative learning in *Aplysia:* Evidence for conditioned fear in an invertebrate. *Science* 211(4481):504–506. [aIG]
Warrington, E. K. (1975) The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology* 27:635–57. [rIG]
White, G., Levy, W. B. & Steward, O. (1990) Spatial overlap between populations of synapses determines the extent of their associative interaction during the induction of long-term potentiation and depression. *Journal of Neurophysiology* 64:1186–98. [RMV]
Wilson, M. A. & McNaughton, B. L. (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* 265:676–79. [BH]
Young, A. W. (1998) *Face and mind.* Oxford University Press. [TS]
Zaidel, D. & Sperry, R. W. (1974) Memory impairment after commissurotomy in man. *Brain* 97:263–72. [DWZ]
Zaidel, D. W. (1990) Memory and spatial cognition following commissurotomy. In: *Handbook of neuropsychology,* ed. F. Boller & J. Grafman. Elsevier. [DWZ]
Zaidel, D. W. & Esiri, M. M. (1996) Hippocampal cell death. *Science* 272:1247–48. [DWZ]
Zaidel, D. W., Esiri, M. M. & Beardsworth, E. D. (1998) Observations on the relationship between verbal explicit and implicit memory and density of neurons in the hippocampus. *Neuropsychologia* 36:1049–62. [DWZ]
Zaidel, D. W., Esiri, M. M. & Harrison, P. J. (1997) Size, shape, and orientation of neurons in the left and right hippocampus: Investigation of normal asymmetries and alterations in schizophrenia. *American Journal of Psychiatry* 154:812–18. [DWZ]
Zaidel, D. W., Oxbury, S. M. & Oxbury, J. M. (1994) Effects of surgery in unilateral medial temporal lobe regions on verbal explicit and implicit memory. *Neuropsychiatry, Neurosychology, and Behavioral Neurology* 7:1–5. [DWZ]
Zeki, S. (1993) *A vision of the brain.* Blackwell. [aIG, JMZ]

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press

https://doi.org/10.1017/S0140525X9926219X Published online by Cambridge University Press