Replies

PATRICIA SMITH CHURCHLAND

Dept. of Philosophy
University of California at San Diego
La Jolla CA 92093
U.S.A.

REPLY TO CORBALLIS

There are three books under consideration in Michael Corballis' review: (1) the book I wrote (2) book he would prefer me to have written, and (3) the book he fears I have written. The dominant theme of the third book is that psychology is codswollop and should be ignored. The second book omits the introductory survey of neuroscience, and focuses instead on experimental psychology, AI models of cognition, and Fodor's modularity hypothesis. Whatever their virtues or flaws, neither of these projects tempted me. Book 3 was unappealing to me because I think it is plainly wrong-headed; Book 2 was unappealing because various incarnations of it are already available. Moreover, it would leave untouched the problem that did interest me greatly, namely whether or not neuroscience can be useful in understanding higher functions. Exploring the *relevance* of neuroscientific research to theories of the mind, therefore, seemed to me both sensible and potentially useful. In what follows I want to amplify briefly these reasons for not writing Books 2 and 3, and perhaps inter alia I can mollify Corballis' feeling that he, and more generally, psychology, have been "slighted" (p. 372).

In the general introduction to *Neurophilosophy* I bare my doctrinal soul: "Thus, the co-evolution of macrotheory and microtheory — broadly, of neuroscience and psychology — is a major methodological theme throughout." (p. 6) I said it, and I meant it. *Co-evolution* of theories at all levels is essential if we are to understand how we see, plan, decide, think, and how we are aware and conscious. Since Corballis has somehow arrived at the opinion that I "reject most of modern psychology" and that I try "to push psychology under the carpet" (p. 364), perhaps it is salutory for me to quote again from the (actual) Introduction:

> So far the ropes thrown across the divide are those from philosophy and from neuroscience, and it will be wondered where ethology and the assorted psychological sciences are thought to fit in the envisaged scheme of things. The fast answer is that they have an absolutely essential role in the enterprise of getting a unified theory of how the mind-brain works. Detailed understanding of the behavioral parameters is essential if we are to know what, exactly, is to be explained by reference to neural

Biology and Philosophy **3** (1988) 393—402.

> mechanisms. Additionally, theories of cognitive and subcognitive processes tendered by psychology, for example, can be expected to co-evolve with neurobiological theories, and these theories are likely to be party to any intertheoretic reduction that eventuates. (p. 9) (See also pp. 373—376.)

The metaphor I offered was this: co-evolutionary progress is rather like two rock climbers making their way up a wide chimney by bracing their feet against the wall, each braced against the back of the other. (p. 374) One of those climbers is experimental psychology, one is neuroscience. It is not the banishment of psychology that I argue for, but the mutual relevance of neuroscience and psychology; it is not the rejection of psychology I seek, but rather the recognition that theory at *all* levels, including the topmost, can be *revised* as the co-evolution of theories proceeds.

To clarify the logical situation here, I should perhaps emphasize one point: my conclusion that neuroscience *is* relevant to theories of mind-brain function by no means entails that psychology *is not* relevant to theories of mind-brain function. The only obvious way to explain Corballis' faith that I reject psychology as worthless is that he interprets my criticism of *folk* psychology as an attack on *scientific* psychology. This is simply a mistake. Criticism of folk physics is not *eo ipso* an attack on physics; criticism of folk biology is not an attack on biology. Quite the reverse. Criticism lays the essential background for discovering a more adequate physics, biology, *and* psychology.

The major reason for espousing the co-evolutionary strategy as opposed to a purely top-down strategy is very simple: one cannot determine the actual nature of information processing procedures solely from input-output parameters. In other words, input-output specifications, however delicate and sensitive, are insufficient to determine either the styles of representation the brain uses or which computational principles are employed. Computational space is consummately vast, and there are indefinitely many procedures that could yield a given input-output profile. We want to know how we *actually* see, plan, learn, and so forth. If we commit ourselves to purely top-down approaches, we deny ourselves important constraints that would help narrow the search space. Ditto, of course, for purely bottom-up approaches.

What is the basis for the conviction, widely shared by cognitive psychologists and philosophers, that theories of higher functions can be explored and discovered quite independently of neuroscience? The principal reason, adverted to by Corballis (p. 365) derives from the claim that cognitive activities are the properties of the cognitive level of organization, where this high level of organization can be likened to a computer program. Just as a given program can be run on diverse machines, so the cognitive program could in principle be implemented in diverse architectures. This shows, the argument goes, that the architecture does not

really matter — what matters is the program. Thus, if you want to explain the behavior of the machine that is running Wordstar, you study the program, not the actual machines it runs on. Those who think the architecture does matter have failed to grasp several simple points: (1) there are three different levels of analysis — the level of the task description, the level of the algorithm, and the level of implementation. (2) the same algorithm can be computed on many different machines.

Insofar as various versions of this argument have motivated widespread ignoring of neuroscience by philosophers and psychologists, I think it worth showing that it is confused in a number of dimensions. To begin with, the claim of independence of the cognitive program from the implementation conflates two very different issues. One concerns whether, as a *matter of discovery*, one can figure out the algorithm and the problem analysis independently of facts about implementation. The other concerns whether, as a *matter of formal theory*, a given algorithm which is *already known* to perform a task in a given machine (e.g., the brain) can be implemented in some other machine that has a distinct architecture.

Now the formal point is straightforward: since an algorithm is formal, no specific physical parameters (e.g., "vacuum tubes", "Ca++") are part of the algorithm. That said, it is important to see that the purely formal point cannot speak to the issue of how best to *discover* the algorithms in fact used by a given machine, nor, for that matter, how best to arrive at a neurobiologically adequate task analysis. Certainly it cannot tell us that the discovery of the algorithms that are actually used by the nervous system will be independent of a detailed understanding of the nervous system. Of course if we already knew how the brain worked, we could set about building machines which used the same algorithms. That much the formal point implies. But we do not know how the brain works, and the formal point is silent on the methodological question of a good research strategy for finding out. (Churchland, Koch and Sejnowski forthcoming)

Moreover, despite the suggestions in the independence argument, algorithms are not indifferent to the architecture. Some algorithms that fall gracefully onto a parallel, analog architecture are handled only slowly and very clumsily by a computer with a serial, digital architecture. Different implementations display enormous differences in speed, efficiency, and elegance, and such considerations will have played a role in the evolution of nervous systems. The matter of time is especially worth emphasizing, because for organisms making a living in the competitive biosphere of hungry predators and unwilling prey, speed and timing are of the essence. Knowledge of brain organization, so far from being irrelevant to the cognitive project, is indispensible for devising likely and powerful algorithms.

Because nervous systems are the product of natural selection, a purely top-down approach to nervous system function can often lead us astray. From the engineering point of view, a specific design for some function, say depth perception, may seem very elegant, yet it may not integrate with

other functions that a nervous system must perform, such as motion perception, thermoregulation, and motor control. And it may not fit very well with the basic structures already in place. The mechanism for the jamming avoidance response in electric fish is a splendid if humbling example of the disparity between a obvious engineering solution and the actual solution Nature found. (Heiligenberg and Rose 1985) The fundamental point is that evolution proceeds by building on structures already in place; it cannot begin from scratch, even though considerations of optimal design might favour this. (Jacob 1982) Consequently, the solutions that evolution stumbled upon may well be less than optimal, and they will be suited to an organism's way of life and environmental niche. The nervous system probably uses assorted tricks to achieve macro effects, and study of the micro-organization of the nervous system will help reveal those tricks (Ramachandran and Anstis 1986).

Finally, it became increasingly clear to me that the singularity conveyed in the descriptions "*the* level of the cognitive program" or "*the* level of implementation" was incorrect. In nervous systems, there is organization and structure at different scales: molecules, membranes, synapses, neurons, nuclei, local circuits, networks, layers, areas and systems. At each of these structurally specified levels, we can raise the computational question: what does this organization of elements do? (Sejnowski and Churchland, 1988) The many levels of structural organization implies that there are also many levels of implementation, each with its own task description and algorithm. Moreover, the same level can be viewed computationally (in terms of its functional role) or implementationally (in terms of the substrate on which the function is implemented), depending on what questions you ask, and on whether you look up or down. From a neuroscientific perspective, some cognitive effects may be the outcome of interactions at the local circuit level (e.g. aspects of early vision such as binocular depth processing), others may reflect processing at the systems level (e.g. the holding of information in short term memory). It is an empirical, not an *a priori* matter, what and how many levels of brain organization figure in explanations of mental phenomena. Neither implementation nor computation defines a single, monolithic level, and hence one of the basic framework assumptions supporting the "irrelevance of neuroscience" doctrine is simply misconceived.

Corballis laments the lack of a comprehensive survey of the psychological literature, as a companion to my discussion of neuroscience. (p. 364) In the Introduction (p. 10), I briefly lamented the lack myself, and explained that the book would have been unapproachably huge had I included such a discussion. Even so, there are a number of places where I do draw on data from experimental and cognitive psychology by way of illustrating the co-evolutionary methodology (e.g., p. 372). Corballis peevishly chides me for using this data on grounds that "these are psychologists, the species that Churchland supposedly disapproves of." (p. 368)

Any author has to make choices concerning what to include, and what must be left out, and judgment may vary rather a lot concerning what examples would serve best. Consistent with the general aims of the project, I was forced to make difficult choices at every turn. For example, *Neurophilosophy* also lacks a history and comprehensive survey of work in theoretical biology, as I noted with regret, (p. 478) but again, this was impossible to include if the book was to be transportable by anything smaller than a wheelbarrow. Almost every reviewer of *Neurophilosophy* has had a favorite list of works, some rather quirky, he thinks it would have been nice for me to have discussed at length. Ideally, I agree; in the real world, I can but demur.

Corballis believes that much of my criticism of the language of thought hypothesis (Fodor 1975) could be deflected if instead of focusing on sentential representation I had addressed *propositional* representation. This is puzzling, since I precisely follow Fodor here in supposing that for central issues to be discussed, the differences between propositions and sentences do not figure. Moreover, propositions are notoriously problematic. If they are not sentences, what are they? They are abstract entities. Well, what kind of abstract entities are they and how can they have a causal role in information processing? The consensus is that the best analysis available is in terms of possible worlds, but this analysis involves serious problems for psychology. For starters, all logical truths get counted as equivalent, and all mathematical truths get counted as equivalent and hence *on this analysis*, belief in one such proposition is indistinguishable from belief in any other. For psychology, that is a disastrous consequence, since clearly I may believe $2 + 2 = 4$ without believing, say, that $n^x \times n^y = n^{x+y}$. And these differences will show up in behavior.

The issues concerning sentences versus propositions as the objects of belief and vehicles of meaning are really very complicated and cannot be sorted out here. Suffice it to say that Corballis merely claims but does not show that any of folk psychology's problems can be dodged by a shift to dubious entities. So far as I can tell, none of the problems I discuss — tacit belief, knowledge access, the frame problem, conflicting criteria for ascription of content, early learning, pattern recognition, representation in nonverbal organisms, representation in sensory modalities — can be solved merely by invoking propositions. I remain convinced that solutions to these problems will require, amongst other things, a systematic look at the brain.

REPLY TO JOHN BISHOP

In John Bishop's careful review, the central question raised is this: am I genuinely an eliminative materialist, or behind the radical fanfare does

there lurk a conservative reductionist? The first option implies that folk psychology, as we currently love and practise it, will turn out to embody a confused and misdirected taxonomy, and to rely on principles and explanatory patterns which fail to fit with what we discover in cognitive neuroscience. On this option therefore, the fate of folk psychology would be much like the fate of Ptolemaic astronomy and the phlogiston theory of combustion. Elegant and useful, perhaps, but in the event, out-manoeuvered in explanatory and predictive virtue by a more powerful scientific framework.

On the second option, it turns out that most of the categories, principles, and explanatory patterns within folk psychology can be smoothly reduced to neurobiological categories, principles, and explanatory patterns. Psychological phenomena, as currently taxonomized, explained, and understood, will be identified with, and thereby reduced to, structures in the nervous system. Should this be how things turn out, folk psychology would share the same status as, say, classical thermodynamics or geometrical optics. Apart from some relatively minor revisions, folk psychology would live on as the core of scientific psychology. Bishop detects some dithering in *Neurophilosophy*, and would prefer to see me quit dodging in and out of the closet. More particularly, he would like to see me emerge as a smooth reductionist.

So which is it to be? My first and most basic response is that is an *empirical* question. *I* do not make that choice. It is a matter of the way the world — or that part of it that consists of nervous systems — truly is. Until we have in place some basic neuroscientific theory concerning core macro phenomena, until we have established some fundamental parts of the story concerning how nervous systems represent and process information, we cannot speak from a firm empirical and theoretical perspective. This is why Paul Churchland and I have on various occasions tried to replace the label "eliminative materialism" with "revisionary materialism". (Churchland, P. S. 1986b) How much revision folk psychology will undergo, whether it will be massive or minor, and whether, assuming it is massive, people forge a new vocabulary or choose to continue with the old words even though their meaning is transmogrified, is not decidable *a priori*. That, as they say, is the bottom line.

Now let's get a bit more speculative. In considering a research program for discovering what one does not yet know, one needs to make some judgments about what strategies will be productive. These judgements are educated guesses, based on the empirical data available so far. For example, if the aim is to understand how the mind-brain works, should we assume the basic integrity and autonomy of folk psychology, and design a program on that assumption? One research program takes that view. It says that a central apparatus for explaining behavior are beliefs and desires. These key elements are representational, minimally in the sense

that they are about things. Since within folk psychology itself, we do not yet have satisfactory explanations for how we reason, plan, decide, remember, see etc. we need to devise a theory that will yield such explanations.

Within the context of this set of problems, Fodor's research program clearly has many virtues. As he sees it, if you expect a theory of information processing to involve complexes of beliefs and desires, then beliefs and desires must be the kinds of things on which mechanisms of information processing can operate. Consequently, those states have to have structure, composition, and logical relations; they have to be amenable to syntactic individuation and manipulation. The only obvious way to do this is to hypothesize that beliefs and desires are relations between persons and sentences in the language of thought. Thus, on this hypothesis, there have to be sentences in the head. It is important to understand that what Fodor was recommending was a research program for working out a genuine theory, in the sense that the theory would describe the mechanisms which would explain how we think, plan and so forth. Moreover, it was a program that could mesh with AI and experimental psychology.

Although this is a research program has many strengths, it also has what Paul Churchland and I take to be flaws of sufficient gravity to question whole project. Thus the catalogue of troublesome aspects of the research program: how can representation in preverbal or nonverbal organisms be explained (the infralinguistic catastrophe), how can we deal with tacit belief, with knowledge access, with the frame problem, with pattern recognition, with nondeductive reasoning, with effective real-time interactions with the motor system? Most pressing, perhaps, how can we understand *learning*? These are not trivial, easily repairable glitches, as the repeated failures to solve them in the conventional AI framework sadly demonstrate. Rather, they motivate the radical suggestion that we ought to look for representations of a kind very different from sentences, and for style of computation and information storage very different from what we see in conventional AI. They motivate the suggestion that the folk psychological basis for the project may itself be so problematic that we have to question whether it is the right basis at all for a scientific psychology.

None of these considerations constitute an *a priori* argument for anything; rather, they are probabilistic arguments based on sizing up the empirical situation as best we can. Thus, the claims concerning radical revision are to be understood as how, given the state of our empirical knowledge and ignorance, I *bet* things will turn out. They are empirically-based arguments for a certain kind of research program. Recognizing that they are empirical arguments, I am also quite willing to say that my bet could be wrong — that folk psychology might turn out to have more

staying power within cognitive neuroscience than it now appears to have. The aforementioned problems (p. 49) might succumb to solutions I never dreamt of. Additionally, while I argue for a research program in which neuroscience is a major component, I am nonetheless unwilling to say that the only research progam to pursue is the one I advocate. It is too early in the game to be sure of very much, and therefore my research *meta-program* is one of tolerance: let many flowers bloom. Thus, philosophers who believe that the Fodor research program will be productive should try to solve the problems it currently faces.

It may be observed that Fodor's research program is not entirely faithful to folk psychology, which, afterall, is not actually committed to anything as specific as sentences in the head. In this spirit, it may be argued that folk psychology needs to be saved from Fodor, and, since many of my objections are directed to Fodor's language of thought hypothesis, saved from me too. The observation about folk psychology may well be correct so far as it goes, but it misses entirely the rationale for what Fodor and a whole generation of cognitive scientists are up to. And what they are up to is getting genuine theories, simulable on a machine, of how cognition works. The point is, we currently understand almost nothing about the nature of human information processing — how we think, see, learn, remember, and so forth. Retreating into the soft arms of a folk psychological framework seems to be a retreat altogether from the project of finding mechanisms and procedures. (See for example, Stalnaker, 1987.)

Alternatives to Fodor's program which, like his but unlike ours, are dedicated to the preservation of folk psychology, are of course welcome. It is not enough, however, to trumpet the longevity, entrenchment, familiarity, and usefulness of folk psychology and leave it at that. Consequent to the trumpeting, we need to have a look at the basic features of that alternative research program. What, even roughly, are they? If not sentences in the head or vectors in the brain, then what?

One could, I suppose, decline to be interested at all in trying to fashion a research program that aims at discovering how the mind-brain works, and opt instead for keeping folk psychology pure, as it were. That, however, is opting out of the scientific enterprise, and I am not much interested in that option. For one thing, it suggests that our cognitive functions are unexplainable, thereby relegating them to the realm of the magical. For another thing, developments in experimental psychology, neuroscience, and computational modeling are converging on theories of information processing in nervous systems, and the possibility that we might really understand aspects of how the mind-brain works is too monumentally exciting to lay quietly aside in favour of caretaking folk psychology.

A central theme in *Neurophilosophy* is that the linguistic model for understanding representations *in general* may need to be abandoned. If we

want to understand how nervous systems represent and process information, one strategy will be to follow in evolution's footsteps, and see how Nature solved the basic problems of perceptual representation, learning, and planning. From that position, we may then be able to understand more complex styles of representing and processing information. In particular, we may be able to address questions about human language. Some representations in mature humans are almost certainly sentential in the way Fodor's program postulates, and some information processing almost certainly involves straightforward deductive inference in the way Fodor's program postulates. But our hypothesis is that substantial parts of learning, perceptual recognition, thinking, planning, deciding, and so forth, do not involve representations and computations of that general kind, and that we need neuroscientific data to inform and inspire hypotheses concerning the nature of these more basic kinds of representation and information processing. This is the research strategy that makes sense to me, and time will tell just how productive or problematic it is.

Finally, Bishop is concerned about whether representational states characterized within cognitive neuroscience would be intentional states. My answer is *no*, in the narrow sense of intentionality that is precisely tied to the linguistic model (i.e. nontruth-functionality, failure of existential generalization, and failure of substitutivity). But my answer is *yes*, in the very loose sense that neurobiological representations are *about* things. And it is this loose sense which I see yielding to a richer explanatory theory of brain-world relations as cognitive neuroscience evolves. Thus I see semantic questions as addressable within the framework of cognitive neuroscience as one element of the wider project of understanding how the mind-brain works.

In assessing the promise of this strategy it is useful to consider that we already know a lot about systematic correlations between response properties in high-level neurons and external stimuli, such as colors, faces, and echo-delays, about neurons responsive to illusory contours in Kaniza figures (von der Heydt, Peterhand and Baumgartner 1984), about neurons with precise reponses to internally generated representational events (Goldman-Rakic 1988), and to attentional factors (Haenny, Maunsell, and Schiller 1988). By virtue of connectionist models, we are beginning to understand what it means for representations to be distributed across a network, for information to be stored in connections of units in a network, and for networks to learn (Hinton 1986; Sejnowski and Rosenberg 1987). But so much more remains to be done.

REFERENCES

Churchland, P. S.: 1986, *Neurophilosophy: Towards a Unified Science of the Mind-Brain*, MIT Press, Cambridge, Mass.

Churchland, P. S.: 1986, 'Replies to Comments', *Inquiry* **29**, 241—272.
Churchland, P. S., Koch, C., and Sejnowski, T. J.: forthcoming, 'What is Computational Neuroscience?' In: *Computational Neuroscience* ed. E. Schwartz. MIT Press, Cambridge, Mass.
Goldman-Rakic, P. S.: 1988, 'Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory', In: *Handbook of Physiology — The Nervous System*, ed. F. Plum and V. Mountcastle, 373—417.
Heiligenberg, W. and G. Rose: 1985, 'Neural Correlates of the Jamming Avoidance Response in the Weakly Electric Fish *Eigenmannia*', *Trends in Neurosciences*, 442—449.
Hinton, G. E.: 1986, 'Learning Distributed Representations of Concepts', *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale NJ, 1—12.
Jacob, F.: 1982, *The Possible and the Actual*, University of Washington Press, Seattle.
Haenny, P. E., J. H. R. Maunsell, and P. Schiller: 1988, 'State-Dependent Activity in Monkey Visual Cortex — 2', *Experimental Brain Research*, **69**, 245—259.
Ramachandran V. S. and S. M. Anstis: 1986, 'The Perception of Apparent Motion', *Scientific American* **254**, 102—102.
Sejnowski, T. J. and C. R. Rosenberg: 1987, 'Parallel Networks that Learn to Pronounce English Text', *Complex Systems* **1**, 145—168.
Sejnowski, T. J. and Churchland, P. S.: 1988, 'Brain and Cognition', In: *Handbook of Cognitive Science*, ed. M. Posner, MIT Press, Cambridge, Mass.
Stalnaker, R.: 1987, *Inquiry*, MIT Press, Cambridge, Mass.
von der Heydt, R., E. Peterhand, and G. Baumgartner: 1984, 'Illusory Contours and Cortical Neuron Responses', *Science* **224**. 1260—62.