PERSPECTIVES ON COGNITIVE SCIENCE

Theories, Experiments and Foundations

edited by

Peter Slezak
University of New South Wales

Terry Caelli
University of Melbourne

Richard Clark
Flinders University

ABLEX PUBLISHING CORPORATION
NORWOOD, NEW JERSEY

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Perspectives on cognitive science : theories, experiments, and foundations / edited by Peter Slezak, Terry Caelli, Richard Clark.
p. cm.
Papers originally presented at the inaugural meeting of the Australasian Society for Cognitive Science, held in Nov. 1990 at the University of New South Wales.
Includes bibliographical references and index.
ISBN: 1-56750-105-2 (cl.).—ISBN: 1-56750-121-4 (ppk.)
1. Cognitive Science—Congresses. 2. Human information processing—Congresses. 3. Cognition—Congresses. I. Slezak, Peter, 1947- . Caelli, Terry. III. Clark, Richard. IV. Australasian Society for Cognitive Science.

BF311.P355 1994
153—dc20

94-36549
CIP

Ablex Publishing Corporation
355 Chestnut Street
Norwood, New Jersey 07648

Contents

TECHNIQUES & APPLICATIONS

FOUNDATIONS

Introduction

Peter Slezak, Terry Caelli and Richard Clark

The papers collected here are representative of the leading work being done in Australia and New Zealand under the banner of 'cognitive science', and the appearance of the volume marks a significant occasion in the development of the interdisciplinary field in this region. Although the present volume has more than parochial value in view of the character and quality of the research reported here, nevertheless, its regional provenance is not without some interest. The papers have been selected from among those which were originally presented at the inaugural meeting of the Australasian Society for Cognitive Science held at the University of New South Wales in November 1990. Coming exactly ten years after the establishment of such a society in the United States in 1980, this conference might be seen, in one sense, as the 'coming of age' of cognitive science in the Australian region. This occasion was the first self-conscious gathering of researchers under the banner of 'cognitive science' in Australia and, in this sense at least, it was a significant step in the direction of a genuine dialogue between scholars in different fields — in a halting pidgin tongue, if not yet in a true interdisciplinary creole. Following the pattern elsewhere, in Australia and New Zealand there are now centers and programs in cognitive science emerging at several universities, and it is hoped that this dialogue will continue to flourish through such centers and through conferences like the one at which these papers were presented.

Of course, these institutional developments come a full 30 years after the establishment of the Center for Cognitive Studies at Harvard in 1960 by G.A. Miller and J. Bruner, and there are grounds for suspecting that the revolutionary developments which swept the United States in the 1960s and 1970s were somewhat attenuated by the time they reached the Antipodes. To take only one significant example, the extraordinary phenomenon of the Chomskian Revolution with its dramatic scientific and institutional impacts has been little in evidence Down Under (see Newmeyer 1986, Gardner 1987). The slower emergence of a truly interdisciplinary cognitive science can perhaps be explained in this way, confirming the remarks of one American psychologist who writes "The extraordinary and traumatic impact of the publication of *Syntactic Structures* by Noam Chomsky in 1957 can hardly be appreciated by one who did not live through this upheaval" (Maclay, 1971, p. 163).

Butterfield, H. (1932). *The Whig interpretation of history.* London: Bell.
Cohen, H.F. (1984). *Quantifying music: The science of music at the first stage of the scientific revolution, 1580-1650.* Dordecht, Germany: D. Reidel.
Collins, H.M. (1984). Concepts and practice of participatory fieldwork. In C. Bell & H. Roberts (Ed.), *Social researching* (pp. 54-64). London: Routledge and Kegan Paul.
Collins, H.M. (1985). *Changing order: Replication and induction in scientific practice,* Beverly Hills, CA: Sage.
Collins, H.M. (1987). Expert systems and the science of knowledge. In W.E. Bijker, T.P. Hughes, & T.T. Pinch (Eds.), *Social construction of technological systems: New directions in the sociology and history of technology* (pp. 329-348). Cambridge, MA: MIT Press.
Collins, H.M. (1990). *Artifical Experts: Social Knowledge and Intelligent Machines,* Cambridge, MA: MIT Press.
Collins, H.M., Green, R.H., & Draper, R.C. (1986). Where's the expertise: Expert systems as a medium of knowledge transfer. In M.J. Merry (Ed.), *Expert systems 85,* (pp. 323-334). Cambridge, UK: Cambridge University Press.
Cooke, N.M., & McDonald, J.E. (1986). A formal methodology for acquiring and representing expert knowledge. *Proceedings of IEEE, 74,* 1422-1430.
Darden, L. (1987). Viewing the history of science as compiled hindsight. *AI Magazine, 8* (2), 33-41.
Debenham, J. K. (1989). *Knowledge systems design.* Sydney: Prentice-Hall.
Feyerabend, P.K. (1975). *Against method: Outline of an anarchistic theory of knowledge.* London: New Left Books.
Freiling, M., Alexander, J., Messick, S., Rehfuss, S. & Shulman, S. (1985). Starting a knowledge engineering project: A step-by-step approach. *AI Magazine, 6* (3), 150-164.
Hall, R.P., & Kibler, D.F. (1985). Differing methodological perspectives in artificial intelligence research. *AI Magazine, 6* (3), 166-178.
Kitakami, H., Kunifuji, S., Miyachi, T., & Furukawa, K. (1984). A methodology for implementation of a knowledge acquisition system. *1984 International Symposium on Logic Programming,* IEEE Computer Society.
Kuhn, T.S. (1962). *The structure of scientific revolutions.* Chicago, IL: Chicago University Press.
Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Cambridge, MA: Harvard University Press.
Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts.* Beverly Hills, CA: Sage Publications.
McCarthy, J. (1958). Programs with common sense. In M. Minsky (Ed.), *Semantic information processing* (pp. 403-410). Cambridge, MA: MIT Press.
McCarthy, J. (1987). Generality in artificial intelligence. *Communications of the ACM, 30,* 1030-1035.
Mettrey, W. (1987). An assessment of tools for building large knowledge-based systems. *AI Magazine, 8* (4), 81-89.
Oxman, R., & Gero, J.S. (1987). Using an expert system for design diagnosis and design synthesis. *Expert Systems, 4,* 4-15.
Plato (1969). *The last days of Socrates* (Trans. by H. Tredennick), Harmondsworth, UK: Penguin.
Polanyi, M. (1967). *The tacit dimension.* London: Routledge and Kegan Paul.
Popper, K.R. (1959). *The logic of scientific discovery.* London: Hutchison.
Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science.* Cambridge, MA: MIT Press.
Ringle, M. (1979). Philosophy and artificial intelligence. In M. Ringle (Ed.), *Philosophical perspectives in artificial intelligence.* Atlantic Highlands, NJ: Humanities Press.
Snow, C.P. (1964). *The two cultures and a second look: An expanded version of the two cultures and the scientific revolution.* Cambridge, UK: Cambridge University Press.
Wittgenstein, L. (1974). *Tractatus logico-philosophicus* (Trans. D.F. Pears & B.F. McGuinness.) London: Routledge and Kegan Paul.

Reduction and Levels of Explanation in Connectionism*

John Sutton

Department of Philosophy
Macquarie University

INTRODUCTION

Recent work in the methodology of connectionist explanation has focused on the notion of levels of explanation. Specific issues in connectionism here intersect with wider areas of debate in the philosophy of psychology and the philosophy of science generally. The issues I raise in this chapter, then, are not unique to cognitive science; but they arise in new and important contexts when connectionism is taken seriously as a model of cognition. The general questions are the relation between levels and the status of levels which have no obvious relation to others. In speaking of levels, what is the connection, if there is one, between explanation and ontology? Which, if any, concept of reduction is applicable to connectionist systems? What kind of legitimacy can the constructs of common sense psychology, or of that version of intentional realism represented by classical symbol-systems AI, have in a full-scale connectionist theory of mind?

In this chapter I address the promising and sophisticated picture of connectionist explanation developed by Andy Clark in his book *Microcognition* (Clark, 1989a) and in a number of recent papers (Clark, 1988a, 1989b, 1990a). The drift is to suggest that, while Clark makes clear the radical nature of the connectionist explanatory framework, his view fails to account successfully for the value of high-level explanations and for why such explanations work. In particular, Clark doesn't provide a sufficiently robust account of the kind of mental causation which seems to be necessary if realism about propositional attitudes is to be maintained. A weak

*Many thanks and my acknowledgements to Daniel Stoljar, with whom I wrote and delivered an ancestor of this paper at the University of Adelaide in March 1990. The outline of A Theory of Reduction and Levels of Explanation presented here was worked out in collaboration with him. Thanks too to George Couvalis and Graham Nerlich for comments on that paper, and to Gerard O'Brien, Huw Price, and Doris McIlwain for many helpful discussions.

requirement of reducibility on a level of explanation, which I will spell out and defend at some length, will explicate the relations between levels in a way Clark's position cannot. It will then serve as a defense of the "condition of causal efficacy" on explanations which Clark, following Jackson and Pettit, rejects. Finally I apply this weak reducibility constraint back specifically to explanation in connectionism, and discuss the status of high-level explanations of connectionist systems.

I will deliberately be blurring some allegedly vital distinctions here. I won't be drawing sharp distinctions between levels of description and levels of explanation, between intertheoretic and interlevel reductions, nor between type and token reductions. The metaphysics of reduction I advance has a number of gaps of detail, but its general drift is so appropriate to connectionist explanation, and such a useful counterweight to Clark's thoroughgoing antireductionism, that its introduction to these specific debates might excuse the compressed form it takes here.

To situate the issue I am addressing, consider the long-running dispute about the legitimacy of causal explanations in terms of propositional attitudes, between robust "big-R" realists and eliminativists (P.M. Churchland, 1981, 1988b, Fodor, 1987). From the perspective of this chapter, the differences between these two positions are relatively minor: The requirement of reducibility on explanation to be expounded is, I claim, weak enough to be accepted by Fodor and by philosophical functionalists as well as by eliminativists. But Clark sees both extremes as misguided (along with others whose specific views I won't be discussing here but who include Dennett, 1978, 1987, 1988; Wilkes, 1984, 1986; Rorty, 1980, 1983; and Price, 1988). Concerning requirements for the legitimacy of common sense psychological explanation, Clark laments,

> We can find two sets of otherwise opposed philosophers united in mutual error. Both Fodor's (1987) defence of ordinary mental talk and various eliminativist attacks on it are committed to a principle of scientific legitimation which can be stated thus:
>
> The goodness of Folk Psychological talk depends on its being legitimised by the discovery of an engineering story which shares its form. (Clark, 1988a, p.275)

It is Clark's rejection of this principle which is open to dispute. The range of reductive possibilities in the development of cognitive science includes, as extremes, Fodorian realism and San Diego eliminativism. But Clark's wish to accept and justify ordinary mental talk and its theoretical derivatives without "scientific legitimation" is, I maintain, neither a genuine option nor a good enough defense.

CLARK'S LIBERALISM ABOUT CAUSATION AND LEVELS

Levels of Connectionist Explanation

Clark offers a neat, innovative model of explanation in connectionist systems (1989a, especially pp. 83-105, 184), and discusses its departures from traditional theories of explanation in cognitive science (1990a). He acknowledges that connectionism simply fails to fit Marr's classical view of levels of explanation (Marr, 1977, 1982). As I am almost wholly in agreement with Clark on these points, I will only sketch, from his presentation, those levels of explanation the status of which make them relevant to my argument.

Starting only with a general task specification, connectionist models with distributed representation are simply set running with random weights, and trained up to better performance, by a variety of supervised or unsupervised learning procedures. The first steps in connectionist explanation, then, are not detailed specifications of the function to be computed and the information on which algorithms are to draw, as in classical models (Peacocke, 1986). Rather they deal with a fully working system. At this stage, the network can be described only by precise mathematical specification of the connections and weights of its individual units. Differential equations can specify both the state of the system at a given time (by stating a vector of numerical values), and its dynamic learning pattern. Explanation of this sort is at what Clark calls "the numerical level" (1989a, pp. 188-189). At this level, as Smolensky notes, "the explanations of behavior are like those traditional in the physical sciences" (1988, p. 1).

Only at this stage can the observer of the connectionist system work backwards, up the levels, toward an understanding of larger-scale patterns of hidden-unit activity, through, for instance, network pathology caused by artificial lesions and hierarchical cluster analysis. I will be concentrating on cluster analysis as an example of an interesting intermediate level of analysis, because it is well-known and already much discussed in the present context.

Cluster analysis reveals the "sorts of internal representations the network has developed to carry out a particular task" (Elman, 1989, p. 5). For each class of input, a mean vector of hidden unit activation patterns is computed. These mean vectors are all then subjected to hierarchical cluster analysis. "The hierarchical interpretation is achieved through the way in which the spatial relations of the representations are organized. Representations which are near one another in representational space form classes, and higher level categories correspond to larger and more general regions of this space" (Elman, 1989, p. 7; 1990, pp. 205-207). Through such analysis the observer can display a hierarchy of partitions, portraying the shape of the representational space of the possible hidden-unit activations that power the network's performance (for more examples and detail see Rosenberg & Sejnowski, 1987; P.S. Churchland & Sejnowski, 1989; P.M. Churchland,

1989a, 1989b; Clark, 1989a, 1990a; Elman, 1989, 1990).

Understanding of connectionist systems, then, comes not through detailed prior analysis of the structure of the task domain, but through statistical glimpses into the running of an already operational network. This "explanatory inversion," which Clark describes as "Marr-through-the-looking-glass" (1990a, p. 215), does not imply a lack of explanatory power. In letting the task organize the network, rather than "imposing the form of our conscious, sentential thought on our models of unconscious processing" (p. 218), the connectionist is able to avoid "ad-hoc organizing principles and sentential, linguistic bias" (p. 219) projected downwards from our conscious understanding[1].

So far so good: Connectionism, in Clark's rendering, is a means to escape what Patricia Churchland called philosophers' "fetishism with respect to logic as the model for inner processes" (1986, p. 381). But what of the status of the levels, discovered in such post hoc strategies as artificial lesioning and cluster analysis, which are higher than that of mere numerical specification of connection weights? There is an intuitive sense in which it is the latter, the base level of connection weights which, as Paul Churchland has put it, drives the dynamic cognitive evolution of the system over time (P.M. Churchland, 1989a, Section 5). What implications does this have for explanation? This question is the focus of some debate, and is central to my concerns in this chapter.

For now I want to bring into play other high levels of description of connectionist systems. Firstly, what Clark calls "the symbolic-AI level": Connectionist systems are described at this level as if they were classical, as if they follow rules, access schemas, fire productions, and so on. These are descriptions according to which the system seems to be satisfying "hard, symbolic constraints in serial order" (1989a, p. 194). The descriptions will tend to break down, to fail to explain, under suboptimal conditions, for "solutions" to ill-posed problems, or with curtailed processing time. In these conditions, connectionist systems will still give "sensible performance" by satisfying as many soft constraints as well as possible. So the formal descriptions of the symbolic AI level cannot provide a unified account of genuinely connectionist cognition, even if they appear to be accurate descriptions of a system's behavior under ideal conditions (see Clark, 1989a, pp. 194-195, and Smolensky, 1988 for detail on this level).

The final relevant level is that of common sense psychology. Here too there is a prima facie tension with the base numerical level. The words and concepts of ordinary language have, it initially seems, no obvious discrete analogs in distributed connectionist systems, since in any particular case they will be

[1] Compare Clark's description of Fodor's search for in-the-head structures which mirror the structures of conscious thought-ascription as "polishing the tip of the iceberg" (Clark 1988b, p. 616).

represented as a complex activation vector across a large set of units in state space. It is this point which drives the recent suggestion that eliminative materialism would be confirmed by the success of connectionism (Stich, 1988a, 1988b; Ramsay, Stich, & Garon, 1990; P.M. Churchland, 1988c, 1989c; disputed by O'Brien, 1991, 1993).

Clark has been at pains to maintain the importance of all these levels of explanation: Cluster analysis, the symbolic AI level and the common-sense psychological level (1989a, pp. 199-201). Surely some at least of these higher levels are going to count as explanatory, even, as Clark says, for the antisententialist unsatisfied with the propositional bias of classical AI (1990a, p. 218). But what principled account can be given of the status of these levels of explanation? It is not with the question of whether such higher level explanations may be justified, but with that of how and why they are justified, and of what would count as a justification, that my disagreements with Clark lie.

Clark's Causal Liberalism

Clark's defence of high-level explanations relies on the grouping of systems into equivalence classes defined according to the purpose of the explanation. "Each such grouping requires a special vocabulary, and the constructs of any given vocabulary are legitimate just insofar as the grouping is interesting and useful" (Clark 1989a, p. 187).

The interest or utility of a level of explanation is, for Clark, sufficient for its legitimacy. In particular, utility is sufficient for the legitimacy of the constructs of any equivalence class even when these constructs are not reducible to the constructs of a physical causal level of explanation. We should in psychological explanation expect no "neat mapping" between ascribed mental states and "scientific stories about the inner causes" of behavior (1989a, p. 57; p. 49, p. 94, p. 112, p. 196; 1988a, p. 267). Individual thoughts, Clark reiterates, "are perfectly real, but they are not the kind of entities that have neat, projectible, computational analogues in the brain" (p. 153).

Clark takes this position partly on the basis of a holistic ascriptivism about mental states: "ascription of a thought is ... ascription of a structured competence within a close-knit semantic domain" (1988a, p. 271; cf. 1988b, 1990b). Beliefs, for instance, are holistically ascribed on the basis of sufficiently rich and flexible bodies of behavior (p. 267). I don't want to quarrel about this ascriptivism: One could consistently accept it while still running the objections I am about to.

My queries are directed, rather, at the second strand of Clark's rejection of "in-the-head realism" (1988a, pp. 267-273). This is his set of views on explanation, causation, and reduction. It is most clearly set out in his rejection of what he calls the "condition of causal efficacy" on psychological explanation. The condition of causal efficacy is as follows: "A psychological

ascription is only warranted if the items it posits have direct analogues in the production (or possible production) of behavior" (Clark 1989a, p. 196).

Clark's stated aim is to provide a picture of high-level explanation dissociated from this condition, one which will give "a more liberal, more plausible, and more useful picture of explanation in cognitive science and daily life" (p. 196). Such a "liberalism about causation" deliberately divorces (not just causal explanation but) causation from reduction (1988a, pp. 272-273).

This distinction between causation and reduction should result, suggests Clark, in a radical division within cognitive science. On the one hand would be an engineering project, seeking the in-the-head causes of behavior. On the other will be what he calls "descriptive cognitive science," which takes "sentential thought-ascriptions as its data and use[s] dynamic, computational models to chart the logical, epistemic or normative relations among thoughts so ascribed" (1988a, p. 274). It gives "a formal theory or model of the structure of the abstract domain of thoughts" (1989a, p. 153). Clark welcomes the prospect of a "rift" resulting from the lack of any "useful relation" between the two kinds of cognitive science (1988a, p. 273, 1989a, p. 159).

This leaves a puzzling lack of clarity concerning the relation between the two kinds of cognitive science, between formal description and engineering story, and more specifically between higher level psychological explanation and physical causal explanation. There are two ways to read Clark on these topics. He could be making what he calls this "distinction between description and cause" in order to deny the causal powers of the constructs of the descriptive project, in particular to reject the implication of mental states in the causation of action. But unless you're a hardened Wittgensteinian, denying mental causation just won't do. Explanation in terms of beliefs, desires, and the rest must be causal explanation of some sort if it is to be legitimate.

Mostly, Clark acknowledges this requirement. His liberalism is, after all, a causal liberalism, and such an extreme brand of neobehaviorism is not his. As he puts it, "the lack of in-head, engineering analogues to individual beliefs and desires need not deprive us of the right to treat beliefs and desires as real, causally active factors in the etiology of human action" (1988a, p. 273). He criticizes Fodor's view that the computational structure of the brain neatly mirrors the descriptive structure of propositional attitude ascriptions: Fodor, Clark says, is guilty of conflating the two kinds of cognitive science in that he adheres to an overstrict model of causation, buttressed by "a fear that beliefs and desires can only be causes if they turn up in formal guise as part of the physical story behind intelligent behavior" (1989a, p. 160; cf 1988a, p. 277, 1988b, p. 609). Clark thinks this fear is groundless:

> All that we need is that there should be some physical, causal story, and that talk of beliefs and desires should make sense of behavior. Such making sense does involve a notion of cause, since beliefs do cause actions. But unless we believe that there is only one model of causation, the physical, this needn't cause any discomfort. (Clark 1989a, p. 160)

What kind of nonphysical model of causation, then, does Clark offer instead? How can truly causal explanation be exempted from the condition of causal efficacy? Abandoning an earlier discussion involving analogies to other allegedly nonphysical cases of causation (1988a; criticized by Tienson, 1990), Clark offers a principled defense of nonphysical causation drawn from Frank Jackson and Philip Pettit (1988; Clark, 1989a, pp. 196-198). Clark applies to connectionist explanation their defense of the view that "features that causally explain need not cause" (Jackson & Pettit, 1988, p. 392). This involves a distinction between "causal process explanations" and "causal programme explanations" (1988, pp. 388-399).

Briefly, a causal process explanation satisfies the condition of causal efficacy in that it cites the actual causally productive features that are efficacious in a particular or given range of cases. These process explanations will be those given by Clark's engineering project. A causal program explanation, on the other hand, cites a feature or property, common across a range of cases, which causally programs the result "without actually figuring in the causal chain leading to an individual action or instance" (Clark, 1989a, p. 198). In explaining a glass vessel's breaking either by its fragility or by an increase in the temperature of the gas inside it (to take two of Jackson and Pettit's examples), we are not citing the particular causes of the shattering, which might be, respectively, the categorical basis of the glass's structure or the impact of a number of molecules on the walls of the vessels (or indeed any of a multitude of ways that fragility or increase in temperature might have been realized) (Jackson & Pettit, 1988, p. 395). Although neither fragility nor increase in temperature causes the breaking, say Jackson and Pettit, they can be said to program the breaking, and thus explain it.

Clark's borrowing from Jackson and Pettit is intended to account for the value of high-level explanations of connectionist networks. The point for Clark is that such program explanations buy us an increase in generality "at the cost of sacrificing the citation of the actual entity implicated in the particular causal chain in question" (1989a, p. 197). It justifies the cluster analytic level of connectionist explanation, which cites global partitions in activation space as its constructs. Networks with distributed representations, when set running with different random distributions of hidden unit connection weights, may turn out to have identical cluster analyses even when embodying entirely different arrangements of individual connection weights (P.M. Churchland, 1989a, Section 5). One cluster analysis, in other words, is multiply realizable at the lower level of numerical specification of dynamic connectivity patterns. Churchland's point that the causal laws of cognitive evolution operate at the level of individual weights rather than at the level

of the partitions in activation space (identified by cluster analysis) can be deflected, Clark thinks, by a good liberalism about causation. The cluster analysis "causally programmes the system's successful performance, but it is not part of any process explanation" (1989a, p. 199).

Clark goes on to give similar justifications for the explanation of connectionist networks at the yet higher levels of both symbolic AI and common sense psychological explanation. Equivalence class groupings at these higher levels may unite what would be otherwise apparently disparate cognitive mechanisms. They are not "mere approximations to the connectionist cognitive truth," but capture constructs which, though themselves causally inefficacious, highlight important facts about an important range of "cognitive constitutions" (1989a, p. 200-201). In other words, the interest or utility of a level of explanation is sufficient for its legitimacy.

Again, it is not the value of these explanations which I necessarily want to question, but rather the explanation of their value. I can agree with Clark that "explanation is a many-leveled thing," and that it is important in cognitive as in other sciences to subsume a single phenomenon "under a panoply of increasingly general explanatory schemas" (1990a, p. 196). In some cases, important higher level similarities between systems realized in different substructs might be invisible at the lower level. In connectionist explanation, in particular, explanations at the level of dynamic activation equations may obscure interesting cognitive similarities which are apparent at a higher level of generalization (Clark, 1989a, pp. 181-182, 197). But, I maintain, Clark's reliance on utility alone as a measure of the legitimacy of a high-level explanation leaves out important detail. The concomitant rejection of the condition of causal efficacy as a necessary condition on a level's legitimacy stems, I propose, from an overstrict idea of the kind of reductionism such a condition entails.

To carry this point, I need to step back for a moment from the specific problems of connectionist explanation, and give a positive account of the relations between explanation, reduction, and causation which will elucidate an acceptably weak constraint of reducibility on a level of explanation. This will then not only justify some high-level explanations, as Clark wants, but give an account of why they work, of the relations between levels of explanation in a way that his causal liberalism cannot. This is inevitably a sketchy treatment of controversial issues in the philosophy of science, detailed treatment of which I pursue elsewhere (Sutton, 1993). I can only plead that the sketchiness is justified by the urgent relevance of these debates to connectionism. They are inspired to some extent by the work on reduction in the psychological context of Richardson (1979), Hooker (1981), Enc (1983), and the Churchlands (P.M. Churchland 1979, 1985; P.S. Churchland 1986, 1988). Most notably, I am going to assume that the supervenience of one level on another entails the reducibility of the supervening level. A number of philosophers have recently argued for this (for example Rosenberg 1985,

Bacon 1986; both criticised by Kincaid, 1987). Not only does the case hold up, but the resulting conception of reducibility is attractively weaker than that accepted in the traditional positivists' arguments for the reductive unification of science. All that is important for my argument against Clark is that something like my picture of reduction and causal explanation both is plausible and promises a robust justification of the legitimacy of high-level explanations.

REDUCTION AND LEVELS OF EXPLANATION

Outline of a Theory of Reduction and Levels of Explanation

Theories and levels are open to reduction if their contents are. (From here, for convenience, I will talk of intertheoretic reduction, but the account applies equally to interlevel reduction). Reduction as an intertheoretic relation is dependent on an ontological relation between theory-contents, the actual things in the world that true theories quantify over, the actual entities, properties, and relations. I'll be talking about property reduction rather than event reduction or any other sort, but the metaphysics is adaptable to most preferred ontologies.

There are two methods of property reduction. First is plain identity. An identity theory simply identifies F-ness, for instance, with G-ness: There are not two properties, but one. Suppose that an identity theory is successful. Then the content of the theories involved, the properties cited in the theories' explanations, turn out to be literally identical. The theories reduce, the two theories are really one, and the two properties are really one.

Of course, the relationship between the properties in question might be more complex than plain identity. The other tool of reduction, besides identity, is supervenience. One difference between them is this: if one property supervenes on another, there remain two properties, whereas if one property is identical with another, to say there are two properties is strictly false.

Another, and perhaps the defining, difference, between identity and supervenience is that they bear different modalities. If F is identical with G, it is impossible that F be identical to H and not to G as well. Whereas, if F supervenes on G, it remains possible that F might supervene on H, and not on G as well. To put the same point differently, F can supervene on many properties, though it is identical with only one.

We can draw out the modality of this point by using a possible worlds analysis. If F is identical to G, it is so in all possible worlds. But if F supervenes on G, it does so in merely some world(s). In a world where F supervenes on G and on no other property, there would be no actual difference between the state of affairs in which F is identical with G, and that in which F supervenes on G.

The difference between these two cases is a matter not of actuality, but

of accompanying modality. For practical purposes, this peculiarly metaphysical, transworld, difference might not be worth worrying about. If you restricted enough the domain of a theory – if, for instance, you restricted it to this world, or to a set of temporal or spatial parts of this world – it would turn out that the class of identical properties and the class of supervenient properties was co-extensive. Of course, the interesting cases are those in which, even within restricted domains, a property supervenes on a whole range of other properties. In these cases, restricting the domain to one particular realization of a supervenient property will often be a pointlessly tedious task: Here the difference between supervenience and identity remains, for practical purposes, marked.

But whatever the details of particular cases, it remains true that properties that supervene can be taxonomized separately, in virtue of the fact that they belong to different transworld classes. This metaphysical analysis gives us a clarification of the notion of levels. One level is different from another not in actuality, but merely in possibility. Levels are distinguished not by the properties they have intrinsically but by the relational properties they bear to other possible worlds. This fact partly serves to explain why levels are so odd.

Compare the case of property reduction by identity. Talk of levels, it seems, is particularly inappropriate. For, if the upper level is identical to the lower level, it is strictly a misnomer to talk of two levels. The only credence the notion of levels can be given in the identity case is a linguistic credence. Identical things can go by different names. For reduction by identity, then, levels are individuated merely linguistically; whereas, for reduction by supervenience, levels are related by their relational properties to other possible worlds.

How then does this metaphysical account of reduction and levels relate to specific problems of explanation in the philosophy of psychology? I suggest that it gives us the materials of a positive alternative to Clark's causal liberalism. His suggestion that the explanatory interest or utility of any equivalence class of systems is sufficient for its legitimacy left puzzles about how high-level explanations can explain, especially causally explain, without citing physical causes. We are owed an account of the relation, whether it is interesting and useful or not, between higher levels of explanation and lower levels.

I suggest, then, that the interest or utility of a level of explanation is neither necessary nor sufficient for the level's legitimacy. Instead, a level of explanation is legitimate if and only if it both:

1. cites real, causally active entities and properties, and
2. is reducible, in the ways specified, to another level of explanation.

Now, Clark's legitimate levels (on some readings at least) do meet the first criterion, since he claims that entities and properties don't have to be physically causally active to be causally active, but I have suggested that more detail needs to be added. That detail is to be found, I suggest, in the spelling out of the second criterion.

I expand first on my negative appraisal of pragmatics in psychological explanation, and discuss the kind of legitimacy that is at issue here, and secondly on how my kind of reductionism should be weak enough to deflate much traditional anti-reductionist criticism. It will then help to reinstate the condition of causal efficacy on explanation, and thus provide robust legitimation for genuine high level explanations.

Reduction, Legitimacy, and Explanatory Utility

The kind of legitimacy for which I have suggested criteria is ontological legitimacy. Explanation, the idea is, cannot just be free-floating: It is, among other things, ontologically committing. This is not intended to be anything like a full-scale theory of explanation. Explanation does a lot of things besides make ontological commitments. In particular, pragmatic considerations will often be of central importance: as van Fraassen suggests (1980), whatever reduces someone's puzzlement can count as an explanation. I am not denying the importance of utility, merely claiming that it isn't all there is to a theory of explanation, deliberately divorcing it from that part of the theory of explanation which arises from taking ontological commitments seriously.

Note first that what is excluded here is the purely epistemological point about what it is for an explanation to be interesting. Many levels of explanation which are legitimate on my criteria will be tedious in the extreme. These will most notably include cases where a higher level explanation cites entities or properties which are defined only functionally, or which are highly disjunctive. Such entities or properties will normally be reducible to a huge disjunction of entities or properties at a lower (micro)level. Examples of this kind would be watches (or the functional state of being a watch), airfoils, crumpled shirts, games, friendships, haircuts, and home runs. In these cases, no pragmatic benefit at all will come from focusing on the lower, reducing, level explanations: Indeed in these cases, as soon as you start to do any reducing at all, there is a likelihood of missing similarities, important to us, which can be understood at higher levels such as the level of description of gross behavior. These things just are realized too variously for the reductive stories to have any pragmatic value. But what you do get from knowledge of the low-level disjunct, or, more commonly, from knowledge that there is a low-level disjunct, is ontological sanction for the high-level construct cited in the explanation. I see no reason why this shouldn't be true even in the large number of cases where the lower level disjunct is open-endedly large (which has been suggested by Fodor (1986a, p. 19), and Kincaid (1987, pp. 344-347), as a problem for reductionism). Inductive confidence that there

is a reduction possible in any particular realizing case is all that the reducibility criterion requires. There really are crumpled shirts, neurons, watches, molecules, haircuts and home runs in a way that there really aren't witches, Ptolemaic epicycles, animal spirits, sunrises, gods, and ghosts. But it seems an advantage of the present account that it leaves entirely open the question of which, of those things which there really are, will be reducible to an explanatorily interesting lower level.

All that this shows is that interest is not necessary for legitimacy: not very controversial. We are all too familiar with the fact that our limited epistemological horizons can't cope with the vast amount of genuine things in the world which could feature in (ontologically) legitimate explanations. To dispute Clark's claim that interest is sufficient for legitimacy, I have to show that the converse point holds, too, in other words, that a level of explanation can be interesting and useful without being legitimate.

Sincere theology and parapsychology, eighteenth century phlogiston theory and animal spirits theory, Cartesian dualism, and the like, are all fascinating discourses. But while the levels of explanation employed in such disciplines may be interesting and useful, they are not (ontologically) legitimate, for they posit entities, properties, and processes which are not reducible in my weak sense, do not exist, and have no causal powers. This is not just a cheap appeal to discredited ontologies: The possibility envisaged by eliminativists (error theorists) about any particular level is exactly that the constructs of that level are of the same status as the "denizens of [these] discredited ontologies" (Dennett, 1988, p. 538) were before they were discredited, that we are now on some critical cusp of conceptual change. The point is that it is impossible to know in advance whether any particular level of explanation is analogous to these eliminated levels or to levels, such as macroscopic levels of geological or meteorological explanation, which are reducible and thus legitimate. You can only find out which of these possibilities holds by looking for possible reductions in the particular case.

It can not, in other words, just be assumed that the psychological constructs cited in the higher levels of psychological explanation have the real causal powers they are thought to. Many of them, or things very like them, probably do exist, and probably do have causal powers pretty much like those which people think they do, but if this is so, it is because they are reducible in my sense to the constructs cited in lower level explanations. Only by way of this reducibility can we understand why such explanations work. If they are legitimate, they are so because of their reducibility, not in spite of their irreducibility. The interest of a level of explanation, then, has little or nothing to do with that level's legitimacy.

Varieties of Reductionism

A common challenge to this kind of reducibility constraint on explanation is that it is, or leads automatically to, a much stronger reductionism. This

stronger reductionism is often described as a kind of a priori view that basic low-level physics is the only serious theory. Reduction, the complaint goes, would entail the elimination of the reduced levels and their replacement by the only legitimate level: Only the entities described in basic physics are "really real." On this view of reduction, there are no "flow" relations between levels – there is only one level. Clark seems to share this vision of what reductionism amounts to, for he claims that his causal liberalism will allay the fear of "the specter of reduction" (1989a, p. 181).

But this picture is misguided. Reduction in the sense outlined above does not entail elimination. This is impossible to stress too strongly, for the assumption that it does – that successful reduction spells the end of the legitimacy of the constructs of the reduced level or theory – still motivates much hostility toward what are actually weaker versions of a reducibility constraint (this misconstrual runs, for example, through the recent critique of physicalism by Crane & Mellor, 1990; see their criticism of Fodor & Field on p. 193 for one instance). But in fact reduction specifically rules out elimination: Successful, smooth reduction, on the contrary, actually guarantees the reality and the legitimacy of the higher reduced level. Finding out what the higher level constructs are identical with or supervenient on tells you what they are, not that they do not exist.

Of course some attempts to reduce may fail: Nothing like a smooth reduction may be possible. In these cases, reductionism may lead to complete displacement of the higher level, but this will be precisely because reduction has failed (P.M. Churchland, 1979, Section 11; P.S. Churchland, 1986, pp. 278-295; Duran, 1988, pp. 296-299). The reducibility constraint allows for a range of reductive possibilities ranging from smooth, retentive reductions, most notably identities, which "preserve ontology" (Hooker, 1981, p. 201) by guaranteeing the legitimacy of the reduced level or theory, to the opposite extreme where there is nothing to which the higher level reduces, and elimination becomes a live option.

So the reducibility constraint is intended to be sufficiently weak to escape criticism of the a priori scientism of that implausibly strong reductionism just sketched. All that it should include is what is naturalistically explicable within the constraints of physicalist monism. All that it should unproblematically exclude are theories which postulate gods, nonphysical minds, "queer" moral values (in Mackie's (1977) sense), phlogiston, animal spirits, and the like. But the ontological legitimizing of psychological explanation doesn't just mean drawing out the implications of materialism. It must also explain the success of psychological explanation. For understanding of how and why it works, there must be demonstrable confidence that the entities and properties it cites are reducible in the specified sense.

It may, however, still be thought that my reducibility constraint is too strong to do the job I want. I was looking for a principled defence of the condition of causal efficacy against Clark's causal liberalism (and other

versions, like Dennett's, of "small-r" realism). My defense is meant to appeal at least to both Fodorian intentional big-R realists and to eliminativists, the two groups Clark sees as "united in [the] mutual error" of the condition of causal efficacy. But then, it may be objected, isn't Fodor in particular notorious for his view that psychology is irreducible to, say, neurophysiology (Fodor, 1981)? And don't functionalists in general support a thesis of the autonomy of psychology from lower level sciences? Haven't I, in other words, misclassified the relevant options on these questions of reduction, causation, and explanation?

I don't think so[2]. That the multiple realizability of psychological states is no bar to a sufficiently weakened reducibility constraint has been argued by a number of reductionists (P.M. Churchland, 1979, p. 112, 1988a; cf. Enc, 1983, Hooker, 1981, Richardson 1979). And here is Fodor in *Psychosemantics:* "It's hard to see ... how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist" (Fodor, 1987, p. 97). The extent to which a Fodorian intentional realist might be a reductionist is captured, I claim, by something very like the weak requirement of reducibility on explanation which I have suggested.

That Fodor could be sympathetic to the kind of account I've sketched is confirmed by his immediately subsequent remarks:

> If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves neither intentional nor semantic. If aboutness is real, it must be really something else. (Fodor, 1987, p. 97)

This is a pretty good statement of the reducibility requirement, and it fits neatly with Fodor's Representational Theory of Mind (RTM; see for example Fodor, 1986b). RTM's vindication of common sense psychology requires "a respectable science whose ontology explicitly acknowledges states that exhibit the sorts of properties that common sense attributes to the [propositional] attitudes" (1987, p. 10). This amounts to a commitment to finding "in-the-head reductive correlates of propositional attitudes," such that the correlates can be straightforwardly implicated in the physical causal chain (Clark, 1988a, p. 268; cf. 1988b).

Fodor, then, specifically accepts the condition of causal efficacy on explanation which Clark rejects[3]. For Clark, like Dennett and Wilkes but

[2] See also Duran (1988, p. 298), discussing extremes of cognitive theories from Pylyshyn to Patricia Churchland: "Virtually no one, so far as I can see, is against the possibility of reduction."

[3] For more on Fodor and Pylyshyn's (weak) reductionism (of classical symbol structures to physical structures in the brain) see Fodor and Pylyshyn (1988), pp. 13-14 and note 9. This is why Dennett has described an "imaginary vindication of the language of thought hypothesis" as "a triumphant cascade through Marr's three levels" (Dennett, 1987, p. 227; cf. Clark, 1990a, pp. 200-203)

unlike Fodor and the Churchlands, legitimate explanations don't have to cite properties which are identical with or supervenient on other properties, or which are directly implicated in the causal process leading to behaviour. What unusually unites the eliminativists with Fodor here is a parallel view about the factors relevant to explanatory legitimacy. They agree that legitimate explanation must cite real, causally active entities and properties, and that these must be reducible, in our weak sense, to another level. Of course the Churchlands make much of the radical consequences of reductionism, because their hunch, contrary to Fodor's, is that the search for reductive correlates of propositional attitudes will generally fail. But it is worth noting again that eliminativism will only be an option to the extent that the reductive enterprise fails.

THE VALUE OF HIGH LEVEL EXPLANATIONS

But what is persuasive about the condition of causal efficacy, even when it is spelled out in these weakly reductive fashions? Are there any considerations which make the requirement of reducibility attractive, other than the odd bedfellows it brings together in Fodor and the Churchlands? I think there are. I'll discuss them first in the specific context of cluster analyses of connectionist systems, then draw some general conclusions about Clark's liberalism, and finally look at the consequences of my approach for the debate over the status of common sense psychological explanation.

Cluster Analysis

Remember the way Clark defended the value of high-level explanations of connectionist networks. The partitions given in a cluster analysis, for instance, play no part in a genuine (low-level) causal process explanation (which would be in terms of connectivity weights alone). But this is no bar to, nor is it relevant to, the legitimacy of cluster analytic explanation, for which, says Clark, its utility is sufficient. In particular, it is a level of explanatory generalization at which are grouped only those systems capable of carrying out a particular task, of satisfying a function in extension. It brings together at a useful level of abstraction all and only the range of networks which can "negotiate that cognitive domain" (Clark, 1989a, p. 199).

Two responses to this are possible in the light of my metaphysical digression. The first is the radical one (originally by Paul Churchland), that genuinely causal explanation will be only at the numerical level, because only such an explanation will account for the specific characteristics of the actual network with respect to learning, generalization, performance on degraded input, and the like (P.M. Churchland, 1989a, Section 5; Clark, 1989a, pp. 193-194). The point would be that only a causal process explanation, and

not a program explanation, will do justice to the phenomena.

But I don't find this eliminativist response satisfying, in the case of cluster analysis anyway.[4] My hunch is that cluster analysis can give us genuine causal process explanations of the results of processing. This second response accepts the value of cluster analytic explanation, but, in contrast to Clark, also has an account of why they are valuable, why they work. They are valuable because they are causal explanations in the strongest, process sense of causal explanation: because they both (a) cite real, causally active entities, properties, and processes; and (b) are (weakly) reducible, in fact supervene on, lower levels. "Causal program" explanations in general must be identical with or supervenient on lower level (process) explanations to be legitimate. The reducibility explains why they are valuable: because the partitions in state space which they cite, for instance, are really there in state space, constituted by, realized by, and supervenient on the (numerical level) connectivity weights in the particular case.

Causal Liberalism Revisited

This idea has already been applied more generally by Mark Rowlands to the examples given by Jackson and Pettit (Rowlands, 1989). While agreeing that functional and disjunctive properties can play a role in true causal explanations, Rowlands notes that this is because they bear some relation to properties which are causally productive (in the "process" sense). Such a property must be realized in an actual case by a lower level property which is causally productive (Rowlands, 1989, p. 272). "The explanatory capacity of the supervenient disjunctive or functional property rides on the causally productive capacity of the property which realizes it" (p. 273). Only the actual realization of an increase in temperature or of a glass's fragility is causally productive in a particular case. I would add to Rowlands' reading that it is only because of such causally active specific realizations that we say the increase in temperature or the fragility causes (and causally explains) the breaking. In each case, increase in temperature or fragility does cause, because both are supervenient on the particular realization in that case.

The point could be extended to bring into question the utility of the causal program/causal process distinction (on which Clark's causal liberalism is

[4] Indeed Churchland later accepts that whether we look to the specific point in weight-space or to partitions in activation space depends, centrally, on what we're doing. He thinks that "while the weights are of essential importance for understanding long-term learning and fundamental conceptual change, the partitions across the activation space, and the prototypical hot-spots they harbor, are much more useful in reckoning the cognitive and behavioral similarities across individuals in the short-term. People react to the world in similar ways not because their underlying weight configurations are closely similar on a synapse-by-synapse comparison, but because their activation spaces are similarly partitioned" (1989b, pp. 234). Utility comes in only here, after the reductive effort, when we already have sophisticated means of relating high level to low level.

based) itself. For antireductionists like Jackson, Pettit, and Clark, seeking to erode fear of "the specter of reduction," the "real" causal processes can surely only occur at the base level of microphysics. For even the numerical levels of connection weights or of synaptic biochemistry are multiply realizable in different configurations of constituent parts: The numerical level stands in the same kind of one – many relation to the subatomic level as does the cluster analytic level to the numerical level itself. To accept the views of causation and explanation espoused by reductionism as the antireductionists construe it would, it seems, lead necessarily to the denial of true causal process explanation at any level above subatomic physics. This, surely, is a *reductio ad absurdum* of their misconstrual of reductionism rather than an accurate picture of a serious view. According to the reductionism they attribute to reductionists, neurophysiological constructs like columnar processing and cell assemblies, or even neurons and their biochemical interactions could not figure in genuine causal process explanation. In Jackson and Pettit's examples, high-level explanation of the glass's breaking in terms of the increase in temperature of the gas inside, or of the glass's fragility, is not in principle any different in causal status to explanation in terms of the impact of molecules on the walls of the vessel (which realizes the increase in temperature), or of the categorical basis of the glass (which realizes its fragility)[5].

Rowlands makes similar observations on the program/process distinction. He thinks that "in itself, the distinction is fundamentally sound:" but this is odd, for he recognizes that "the use of program explanations in science is very widespread indeed ... natural science must deal almost exclusively in program explanations." If this is so, how can the distinction be sound, when one side of it is all but empty? "[A]lmost all the explanatory properties invoked by even a foundational science such as physics" are cited only in program explanations (Rowlands, 1989, p. 271). But apart from showing the emptiness of the distinction he claims to support, Rowlands also reiterates the traditional view that multiple realizability debars reduction: To deny the ubiquity of program explanations even in science is to fall victim to what Blackburn calls the Tractarian View of physical properties: The mistake of supposing that for any physical property there should be a story, in terms of the configuration of some constituent things, saying what it is. (Blackburn, 1991, pp. 206-208).

[5] One response here would be to refer to the view of David Braddon-Mitchell and John Fitzpatrick that true causation does occur only at the microstructural level. Only the actual microstructural instantiation of a high level does the causing, and high level regularities merely explain (Braddon-Mitchell and Fitzpatrick, 1990, section 4). I, in contrast, want to maintain true high level causation where suitable reductive relations between levels hold. Braddon-Mitchell and Fitzpatrick tend towards the neo-behaviourist position that the implication of mental states in the causation of action is an unnecessary requirement. Whatever the merits of their view, it won't help Clark, for he is committed to genuine causation at the higher levels of explanation.

The case is supported by the notorious example of the multiple realizability of temperature across different physical bases for solids, gases, plasma, and vacua. But the disjunctive nature of the physical bases is no objection to particular (domain-restricted) reductions. The Tractarian View can be usefully modified and, contra Blackburn, supported by adding the (later) Wittgensteinian point that a family resemblance is all that is required among reductive realizations. There are many possible but somehow related stories, in terms of configurations of constituents, to be told about your average high-level physical property (for weakly reductionist treatments of the temperature case, see Hooker, 1981, pp. 47-49; P.M. Churchland, 1988a, Ch. 2).

Common Sense Psychology and the Condition of Causal Efficacy

Folk-psychological explanation is, for Clark, "just one more layer in rings of ever-more explanatory virtue" (1989a, p. 200). Like other high-level explanations, it groups "apparently disparate physical mechanisms into classes that reflect our particular interests" (p. 201). But the utility of common sense ascriptions of mental states on the basis of behavior, for Clark, implies nothing about in-head processing, about the nature of the "engineering" account of the causes of that behavior (1988a, 1988b). Folk psychology works as a descriptive model which fixes on important regularities in behavior, not because there are any reductive correlates of its explanatory constructs in the head.

But such a defense of common sense psychology against the advance of neurophilosophy is unilluminating: it simply leaves unexplained the relation between the descriptive account and the engineering account. Clark does assert that his criterion of interest for the legitimacy of a level should not be taken to allow that "anything goes" (1989a, p. 201), but he still gives no principled grounds other than interest for distinguishing legitimate from illegitimate explanations. This defect is perhaps partly to be remedied by his commitment to mixed models of cognition, where connectionist and language of thought systems are working in tandem (1989a, Ch.7; 1989c, 1990b). In these cases some aspects of the descriptive project will, presumably, be backed by an engineering, in-the-head story which shares its form. But to the extent that Clark does support the connectionist rejection of causal processing descriptions which mirror the form of natural language semantics, his overview of causal explanation doesn't tell us why common sense explanation works as well as it does.

Even with functional concepts, defined in terms of a causal/functional role and not in terms of the occupant(s) of the role, we still need an account of how all the particular occupants come to fill the role, of what it is that makes them the sorts of things which can fill the role. If you're a functionalist about watches, or about beliefs, you still need a story of how specific

physical configurations of lower level entities come to play the role watches or beliefs play. In the watch case, the reductive story will be very tediously disjunctive. Watches are so variously realized that, given our inductive confidence that reducibility could go through in any particular case, we won't tend to do it because it won't be very interesting. But we just don't know yet what reductive stories about mental states and processes would be like, where on the continuous spectrum of reductive possibilities they would fit. Would they reduce smoothly, to the retentive extreme (as suggested by both Fodor's Representational Theory of Mind and O'Brien's connectionist vindication of folk psychology)? Would they prove entirely irreducible, as forecast on the Churchlands' eliminative extreme? Or would they fall somewhere in between (as suggested, for example, by Smolensky's (1988, pp. 59-61) limitivism, which sees high-level cognitive explanation as approximately correct, as falling in the middle of the range of reductive possibilities for high-level explanation).

Tienson, criticizing Clark, has noted that even if mental states are individuated functionally or conceptually, the fact that something satisfies this conceptual demand on a kind of ascription is still an empirical, not a conceptual fact, and requires empirical explanation. The conceptual/functional demand does not, as Tienson puts it, explain its own satisfaction. "Quite the opposite. A satisfied a priori demand requires an empirical explanation. That it be to a considerable degree liquid is a conceptual demand on calling something soup. But its being liquid is an empirical fact, subject to empirical explanation" (Tienson, 1990, p. 160). Without a theory relating causal explanation at different levels to the reductive relations between those levels there seems little prospect of such empirical explanation.

Because Clark is wary of the ontological commitment of explanation, he cannot account for the way genuine true explanations latch onto the world. Explanations must, mostly at least, cite real entities, properties, and processes. If they did not, they wouldn't tend to work as often. But Clark, by exempting high-level constructs from the condition of causal efficacy and the need for (weak) reducibility, leaves them ontologically loose and free-floating. For him, the explanatory utility of high-level explanation is all that is required. But this is to close off a priori the discovery of error and of any possibility of revision of the high level. This attitude is perhaps clearest where Clark is discussing Fodor's attempt to find computational structure in the brain which reductively mirrors the structure of propositional attitudes:

> Fodor's approach is dangerous. By accepting the bogus challenge to produce syntactic brain analogues to linguistic ascriptions of belief contents, he opens the Pandora's box of eliminative materialism. For if such analogues are not found, he must conclude that there are no beliefs and desires. The mere possibility of such a conclusion is surely an effective *reductio ad absurdum* of any theory that gives it house space. (Clark, 1989a, p. 160)

This is a transcendental argument against eliminativism. If any theory allows the possibility that its own falsity would entail eliminativism's truth, that theory must be ruled out a priori as absurd. This is a strange argument, to say the least. Transcendental arguments against eliminativism are at best fairly pointless, since they have no prospect of ever convincing those against whom they are aimed, and at worst seriously misguided. Eliminativism may be implausible, but it is not incoherent (see Devitt's 1990 critique of Boghossian, 1990, for criticism of another such transcendental argument). The spectrum between "big-R" realism and eliminativism about the theoretical analogs of the concepts of common sense psychology is an exhaustive one. To refuse a position on it is to enshrine our present common sense as a priori true, and this could reasonably be considered not so much nicely liberal as dangerously conservative. The polemical point of early eliminativism was to erode the air of "a priori sanctity" (P.M. Churchland, 1982, p. 231) around folk psychology. The folk require a more robust defense than Clark can give them: The price of realism about common sense psychology is the requirement to produce empirical, not conceptual, refutations of eliminativism.

REFERENCES

Bacon, J. (1986). Supervenience, necessary coextension, and reducibility. *Philosophical Studies, 49,* 163-176.

Blackburn, S. (1991). Losing your mind: Physics, identity, and folk burglar prevention. In J. Greenwood (Ed.), *The Future of Folk Psychology.* Cambridge, UK: Cambridge University Press.

Boghossian, R. (1990). The status of content. *Philosophical Review, 99,* 153-184.

Braddon-Mitchell, D., & Fitzgerald, J. (1990). Explanation and the language of thought. *Synthese, 83,* 3-29.

Churchland, P.M. (1979). *Scientific realism and the plasticity of mind.* Cambridge, UK: Cambridge University Press.

Churchland, P.M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78,* 67-90.

Churchland, P.M. (1982). Is "thinker" a natural kind? *Dialogue, 21,* 223-238.

Churchland, P.M. (1985). Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy, 82,* 8-28.

Churchland, P.M. (1988a). *Matter and consciousness* (Second ed.). Cambridge, MA: MIT Press.

Churchland, P.M. (1988b). Folk psychology and the explanation of human behavior. *Proceedings of the Aristotelian Society, 62,* 209-221.

Churchland, P.M. (1988c). The ontological status of intentional states: Nailing folk psychology to its perch. *Behavioral and Brain Sciences, 11,* 507-508.

Churchland, P.M. (1989a). On the nature of theories: A neurocomputational perspective. In C.W. Savage (Ed.), *On the nature of theories* (Minnesota Studies in the Philosophy of Science, Vol. 14).

Churchland, P.M. (1989b). Learning and conceptual change. In P.M. Churchland, *A neurocomputational perspective: The nature of mind and the structure of science.* Cambridge, MA: MIT Press.

Churchland, P.M. (1989c). *A neurocomputational perspective: The nature of mind and the structure of science.* Cambridge, MA: MIT Press.

Churchland, P.S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain.* Cambridge, MA: MIT Press.

Churchland, P.S. (1988). Reduction and the neurobiological basis of consciousness. In A.J. Marcel & E.J. Bisiach (Eds.), *Consciousness in contemporary science.* Oxford, UK: Clarendon Press.

Churchland, P.S., & Sejnowski, T.J. (1989). Neural representation and neural computation. In L. Nadel et al. (Eds.), *Neural connections, mental computation.* Cambridge, MA: MIT Press.

Clark, A. (1988a). Thoughts, sentences, and cognitive science. *Philosophical Psychology, 1,* 263-278.

Clark, A. (1988b). Critical notice: Psychosemantics. *Mind, 97,* 605-617.

Clark, A. (1989a). *Microcognition: Philosophy, cognitive science, and parallel distributed processing.* Cambridge, MA: MIT Press.

Clark, A. (1989b). Beyond eliminativism. *Mind and Language, 4,* 251-279.

Clark, A. (1989c). *Connectionism, non-conceptual content, and representational redescription* (Cog. Sci, Research paper CSRP 143). Sussex, UK: University of Sussex.

Clark, A. (1990a). Connectionism, competence, and explanation. *British Journal for the Philosophy of Science, 41,* 195-222.

Clark, A. (1990b). Belief, opinion, and consciousness. *Philosophical Psychology, 3,* 139-154.

Crane, T.M., & Mellor, D.H. (1990). There is no question of physicalism. *Mind, 99,* 185-206.

Dennett, D.C. (1978). *Brainstorms.* Montgomery, VT: Bradford Books.

Dennett, D.C. (1987). *The intentional stance.* Cambridge, MA: MIT Press.

Dennett, D.C. (1988). Precis of *The intentional stance. Behavioral and Brain Sciences, 11,* 495-546.

Devitt, M. (1990). *Transcendentalism about content.* Paper presented at the AAP Conference, Sydney, Australia, July 1990.

Duran, J. (1988). Reductionism and the naturalization of epistemology. *Dialectica, 42,* 295-306.

Elman, J. (1989). *Representation and structure in connectionist models* (Tech. Rep. CRL-8903). San Diego, CA: University of California, San Diego, Center for Research in Language.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Enc, B. (1983). In defense of the identity theory. *Journal of Philosophy, 80,* 279-298.

Fodor, J.A. (1981). *Representations.* Cambridge, MA: MIT Press.

Fodor, J.A. (1986a). Why paramecia don't have mental representations. *Midwest Studies in Philosophy, 10,* 3-23.

Fodor, J.A. (1986b). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind, 95,* 76-100.

Fodor, J.A. (1987). *Psychosemantics.* Cambridge, MA: MIT Press.

Fodor, J.A. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition, 28, 3-71.

Hooker, C.A. (1981). Towards a general theory of reduction. *Dialogue, 20,* 38-59, 201-236, 496-529.

Jackson, F., & Pettit, P. (1988). Functionalism and broad content. *Mind, 97,* 381-400.

Kincaid, H. (1987). Supervenience doesn't entail reducibility. *Southern Journal of Philosophy, 25,* 343-356.

Mackie, J. (1977). *Ethics: Inventing right and wrong.* Harmondsworth, UK: Penguin.

Marr, D. (1977). Artificial intelligence: A personal view. In J. Haugeland (Ed.), *Mind design.* Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision,* Cambridge, MA: MIT Press.

O'Brien, G. (1991). Is connectionism common sense? *Philosophical Psychology, 4,* 165-178.

O'Brien, G. (1993). The connectionist vindication of folk psychology. In S. Christensen & D. Turner (Eds.), *Folk psychology.* Hillsdale, NJ: Erlbaum.

Peacocke, C. (1986). Explanation in computational psychology: Language, perception and level. *Mind and Language, 1,* 101-123.
Price, H. (1988). *Facts and the function of truth.* Oxford, UK: Blackwell.
Ramsey, W., Stich, S.P., & Garon, J. (1990). Connectionism, eliminativism, and the future of folk psychology. In J. Tomberlin (Ed.), *Philosophical Perspectives,* Vol. 4, Atascadero, California: Ridgeview Press.
Richardson, R.C. (1979). Functionalism and Reductionism. *Philosophy of Science, 46,* 533-558.
Rorty, R. (1980). *Philosophy and the mirror of nature.* Princeton, NJ: Princeton University Press.
Rorty, R. (1983). Method and morality. In P. Rabinow et al. (Eds.), *Social science as moral inquiry.* New York: Columbia University Press.
Rosenberg, A. (1985). *The structure of biological science.* Cambridge, UK: Cambridge University Press.
Rosenberg, C.R., & Sejnowski, T.J. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1,* 145-168.
Rowlands, M. (1989). Discussion of Jackson and Pettit, "Functionalism and broad content". *Mind, 98,* 269-275.
Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11,* 1-73.
Stich, S.P. (1988a). From connectionism to eliminativism. *Behavioral and Brain Sciences, 11,* 53-54.
Stich, S.P. (1988b). Connectionism, Realism and realism. *Behavioral and Brain Sciences, 11,* 531-532.
Sutton, J. (1993). *Connecting memory traces: studies of neurophilosophical theories of memory, mental representation, and personal identity from descartes to new connectionism.* Doctoral dissertation, University of Sydney, Australia.
Tienson, J. (1990). Is this any way to be a realist? *Philosophical Psychology, 3,* 155-164.
van Fraassen, B. (1980). *The scientific image.* Oxford, UK: Oxford University Press.
Wilkes, K.V. (1984). Pragmatics in science and theory in common sense. *Inquiry, 27,* 339-361.
Wilkes, K.V. (1986). Nemo psychologicus nisi physiologicus. *Inquiry, 29,* 165-185.

Author Index

Page numbers in *italics* indicate figures