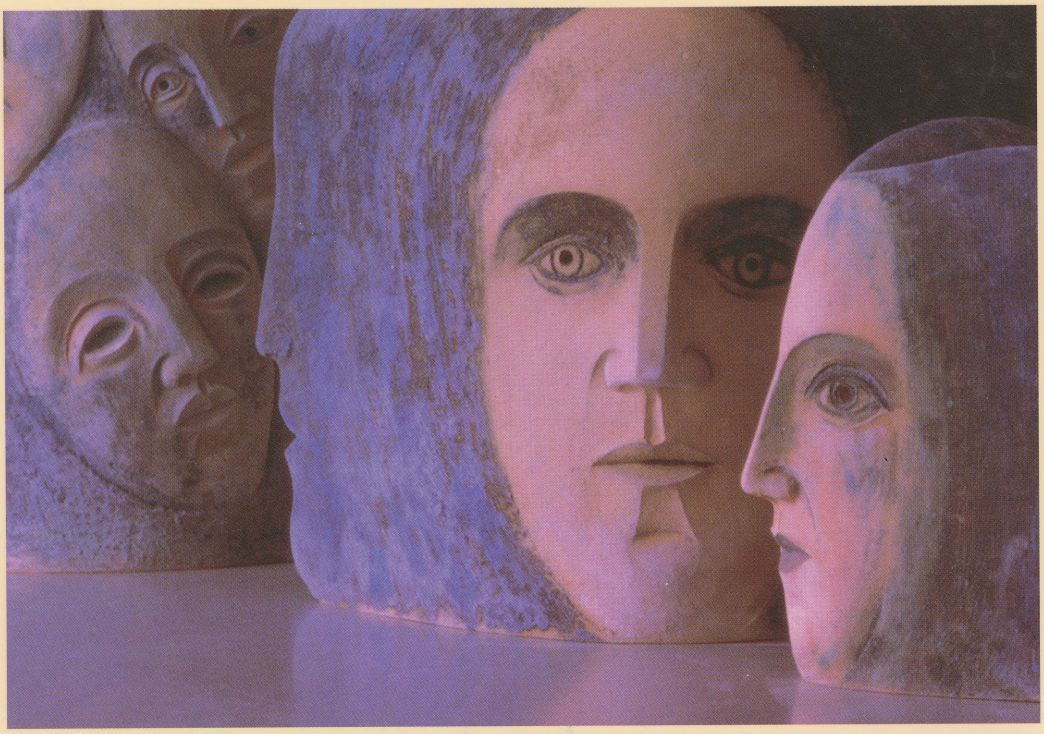
Blackwell Companions to Philosophy

# A Companion to Cognitive Science



*Edited by*

WILLIAM BECHTEL AND GEORGE GRAHAM

B

*Blackwell*
*Companions to*
*Philosophy*

# A Companion to Cognitive Science

*Edited by*

WILLIAM BECHTEL and
GEORGE GRAHAM

*Advisory editors*

David A. Balota
Paul G. Chapin
Michael J. Friedlander
Janet L. Kolodner

BLACKWELL
Publishers

1998

Copyright © Blackwell Publishers Ltd, 1998

First published 1998

2 4 6 8 10 9 7 5 3 1

Blackwell Publishers Inc.
350 Main Street
Malden, Massachusetts 02148
USA

Blackwell Publishers Ltd
108 Cowley Road
Oxford OX4 1JF
UK

All rights reserved. Except for the quotation of short passages for the purposes of criticism and review, no part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher.

Except in the United States of America, this book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser.

*Library of Congress Cataloging-in-Publication Data*

A companion to cognitive science / edited by William Bechtel and George Graham; advisory editors, David A. Balota . . . [et al.].
p. cm. — (Blackwell companions to philosophy; 13)
Includes bibliographical references and indexes.
ISBN 1–55786–542–6 (hardcover: alk. paper)
1. Cognitive science. I. Bechtel, William. II. Graham, George, 1945– . III. Balota, D. A. IV. Series.
BF311.C578 1998
153—dc21 97–38757
CIP

*British Library Cataloguing in Publication Data*

A CIP catalogue record for this book is available from the British Library.

Typeset in 10 on 12 pt Photina
by Graphicraft Typesetters Limited, Hong Kong
Printed in Great Britain by T.J. International, Padstow, Cornwall

This book is printed on acid-free paper

# 48

# Levels of explanation and cognitive architectures

ROBERT N. MCCAULEY

## Introduction

Some controversies in cognitive science, such as arguments about whether classical or distributed connectionist architectures best model the human cognitive system, reenact long-standing debates in the philosophy of science. For millennia, philosophers have pondered whether mentality can submit to scientific explanation generally and to physical explanation particularly. Recently, positive answers have gained popularity. The question remains, though, as to the analytical level at which mentality is best explained. Is there a level of analysis that is peculiarly appropriate to the explanation of either consciousness or mental contents? Are human consciousness, cognition, and conduct best understood in terms of talk about neurons and networks or schemas and scripts or intentions and inferences? If our best accounts make no appeal to our hopes or beliefs or desires, how do we square *those* views with our conception of ourselves as rational beings? Moreover, can models of *physical* processes explain our *mental* lives? Does mentality require a special level of rational or cognitive explanation, or is it best understood in terms of overall brain functioning or neuronal or molecular or even quantum activities – or any of a dozen levels of physical explanation in between? Also, regardless of how they compare with explanations cast at physical levels, what is the status of psychological explanations that appeal fundamentally to mental contents?

As a means for beginning to address such questions, proposals about cognitive architecture concern which kind of explanation best characterizes primitive psychological activities. Although, technically, approaches to modeling those activities are unlimited, two strategies have enjoyed most of the attention. The prominence of the classical account and the distributed connectionist (or parallel distributed processing (PDP)) account, notwithstanding, nothing bars the development of other proposals.

Classicism employs rules that apply to symbolic representations to explain cognitive processing. PDP systems propagate activation through networks of processing units from an input layer to an output layer, without appealing to either symbols or their (rule-governed) manipulation. Proponents of these views debate whether a PDP account of cognition characterizes the human cognitive architecture or merely supplies details concerning the implementation of a classical architecture. Their answers depend upon how they regard the notion of cognitive architecture, how they assess the adequacy and centrality of classical accounts, and how they interpret PDP models. Their answers also depend upon what they assume about the relationships between scientific inquiries aimed at explaining the same phenomena but proceeding at different explanatory levels.

The second section discusses analytical levels in science and surveys philosophical accounts of reductionism. The third section considers the question of our cognitive architecture, outlines the classical account and the challenges it poses to connectionism, and surveys various connectionist responses. The final section describes recent integrative models of cross-scientific relations and their implications for these discussions.

## Levels of explanation and inter-theoretic reduction

Scientists' facility with the concept *explanatory level* notwithstanding, clear, unambiguous criteria exist neither for specifying the notion of an explanatory (or analytical) level nor, often, even for distinguishing particular levels. *Within the cognitive sciences* computer scientists use these terms to describe hierarchies of compiled programming languages. Philosophers of science, by contrast, use them to talk *about the cognitive sciences'* relations to one another. This second use (which, arguably, encompasses the first) is especially helpful when considering the bearing of levels of explanation on hypotheses about cognitive architecture arising from *multidisciplinary* enterprises in cognitive science.

Many criteria for locating levels of explanation *among the sciences* roughly converge – at least with respect to theorizing about the structural relations of systems. For example, analytical levels partially depend upon viewing nature as organized into *parts and wholes* and largely mimic *levels of aggregation* (as opposed to simple considerations of scale). If one entity contains others as its parts, and its description requires further organizing principles beyond those concerned with those parts, then it occurs at a higher level of aggregation. The *range* of the entities that constitute any science's primary objects of study and that science's principal units of analysis also track this arrangement of analytical levels. The lower a science's analytical level, the more ubiquitous the entities it studies. For example, subatomic particles, discussed in physics, are the building blocks of all other physical systems (molecules, biological systems, galaxies, social groups, and more). Although complexity has no simple or single measure, the relative complexity of the (aggregated) systems generates a similar picture. Sciences at lower analytical levels study (relatively) simpler systems, at least to the extent that increasingly higher-level sciences deal with increasingly restricted ranges of events concerning increasingly organized systems whose study requires additional explanatory principles. The order of analytical levels also corresponds to the chronological *order in natural history* of the evolution of systems. The lower a science's level, the longer the systems it specializes on have been around.

Presumably, our most successful theories provide important clues about the furniture of the universe. This suggests that levels of analysis in science correspond to levels of organization in nature. Typically, what counts as an entity depends on both the redundancy of spatially coincident boundaries for assorted properties and the common fate (under some *causal* description) of the phenomena within those boundaries. So, for example, both their input and output connections and their various susceptibilities to stains aid in identifying cortical layers in the brain (see Article 4, BRAIN MAPPING). Emphasizing causal relations insures that the sciences dominate such deliberations. The greater the number of theoretical quarters from which these ontological commitments receive empirical support, the less troublesome is the circularity underlying an appeal to *levels of organization* as criteria for their corresponding levels of analysis.

Methodological considerations also segregate analytical levels, but less systematically. Sciences at different analytical levels ask different questions, promote different theories, and employ different tools, methods, and standards. Theories at alternative explanatory levels embody disparate idealizations that highlight diverse features of the phenomena. Such criteria can serve to arrange the major scientific families into levels (physical, biological, psychological, and sociocultural sciences), but each of these families includes separate sciences that, in turn, contain sublevels. We can identify seven sublevels within neuroscience alone (molecules, synapses, neurons, networks, maps, subsystems, and the central nervous system overall).

When analyzing inter-level relations in science, philosophers have cut through these vaguenesses surrounding the identification and differentiation of explanatory levels. They have concentrated on only one relation (reduction) between only one component of levels (theories). Traditional reductionism conceives of all inter-theoretic relations as explanatory and of all explanations as deductive inferences in which at least one of a theory's laws serves as a premise. In the case of reductive explanation, the immediate goal is to show that, with the aid of bridge laws, the explanatory principles of successful upper-level theories follow as deductive consequences from the laws of lower-level theories. The bridge laws establish connections between the two theories' predicates, providing grounds for the explanation of the upper-level theory and for the revelation that its entities are *nothing but* combinations of lower-level entities. Thus, in principle at least, the lower-level theory can, allegedly, replace the upper-level theory without explanatory or ontological loss.

Arguments persist about every feature of this proposal, but those about the bridge laws' status matter most. Ambitious reductionists, who aspire to both explanatory consolidation and ontological economies, argue that only comprehensive inter-theoretic *identities* of the two theories' predicates will insure the desired results. Type-identity theorists find ambitious reductionism particularly congenial, since that view claims that a successful reduction of psychology to neuroscience will certify the identity of mind and brain.

Securing a reductive explanation on the traditional view, however, involves no more than the bridge laws specifying lower-level conditions *sufficient* for upper-level patterns. Under those circumstances, reductions prove domain-specific, limited in scope, and less sweeping ontologically, since any systematic bridge laws will probably apply only in circumscribed settings. For example, even the reductive explanation of the Boyle–Charles law by statistical mechanics reduces not the notion of temperature but only *temperature of a gas*. For a variety of reasons, most philosophers are not optimistic about the possibility of obtaining comprehensive identities between psychological and neuroscientific predicates. This less ambitious view of reductive explanation recommends detailed analyses of scientific research on the relevant systems. Inevitably, the complexities which such analyses reveal do not readily lend themselves to either easy or comprehensive ontological pronouncements.

Two other reactions to the projected failures of systematic inter-theoretic mappings between psychology and neuroscience have gained attention. The first is *eliminativism*, which seeks the same economies as do the ambitious reductionists but does so by exploiting another dimension of traditional reductionism. Defenders of the traditional account regarded it not merely as a model of inter-theoretic relations between different analytical levels, but also as an account of theoretical *progress* within a particular

level. They especially emphasize the corrections which a successor theory offers its predecessor. Noting that in some cases of intra-level theoretical progress the requisite bridge principles – let alone inter-theoretic identities – were not even remotely plausible, critics of traditional reductionism spotlighted episodes in which victorious theories simply eliminated their predecessors. The oxygen theory of combustion did not reduce the phlogiston theory; it eradicated it. Such episodes illustrate the most extreme form of inter-theoretic correction. Eliminativists in the philosophy of psychology apply these lessons to the inter-level case of neuroscience and psychology, holding that if psychological theories do not map reasonably well onto neuroscientific theories, then they will undergo elimination, in light of the neurosciences' superior merits and promise (Churchland, 1989).

Traditional reductionists and eliminativists conflate disparate forms of inter-theoretic relations when applying the same model of reduction both to (1) relations between theories at different explanatory levels and (2) theoretical progress within a single explanatory level (McCauley, 1986). Inter-theoretic corrections can occur in both sorts of cases; they must in the second. Elimination often occurs in the second too. But at least in the science of the past century elimination is virtually nonexistent in the first – certainly *when both* the upper-level science (experimental psychology in this case) is institutionally well established and the elimination is alleged to span the divisions between the major families of levels listed on page 612 (as the elimination of psychology by neuroscience is). (This contrasts with merely *consolidating* a theoretical account of what had previously been regarded as diverse phenomena at various sublevels *within a single level* – in the way, for example, that Maxwell's theory of electromagnetism did.)

The other prominent response to reductionism also questions the availability of adequate inter-theoretic connections. Jerry Fodor defends the *irreducibility of psychology* by insisting on the letter of ambitious reductionism. If bridge laws must provide type identities between psychological and neural predicates, then, Fodor (1975) argues, successful reduction will prove virtually impossible. Fodor does not deny that psychological states are brain states. He just denies the availability of *systematic* connections capable of linking theoretical predicates. He does not repudiate the identity of psychological and neural *tokens*; he just rejects the identity of psychological and neural *types*. Each token of some psychological type is a token of some physical type; however, every token of that psychological type is not a token of *one* particular physical type.

Two general considerations encourage Fodor's skepticism. The first concerns the disparate explanatory tactics pursued at different analytical levels. Psychology and neuroscience manifest all the methodological dissimilarities between analytical levels outlined on page 612. They often study human cognition and behavior in radically different ways. To the extent that psychological – but *not* neuroscientific – investigations presuppose conceptions of rationality, these two disciplines address distinct concerns and utilize idealizations that diverge drastically sometimes. Consequently, they often spawn explanatory principles and predicates that interpret closely related phenomena in substantially different ways. Thus, Fodor argues that psychology and the other *special sciences* formulate generalizations concerning types whose tokens' physical descriptions frequently have little or nothing in common. Fodor notes, for example, the diversity of physical things that serve as money (let alone those that might instantiate some belief). Such diversity blocks inter-level identities, because types of states and processes

construed psychologically correspond only to large disjunctions of types construed physically. On this view, psychological states are so multiply realizable in our neural substrate that any possible connecting principles would prove neither theoretically interesting nor heuristically useful, since they would contain unmanageably large disjunctions of predicates.

The second ground for skepticism about securing adequate bridge laws suggests an even broader sort of multiple realizability and directly links questions about how analytical levels relate to questions concerning our cognitive architecture. Workable bridge laws are particularly unlikely when, as in psychology (of both the commonsense and the scientific varieties), the complex systems under study demand theories that often characterize states and processes *functionally*. Psychological theorizing has bred far clearer accounts of what such things as desires, parsers, mental images, and episodic memories *do* than of what they *are*.

Our inclination to assign such capacities and states not only to other animals, but especially to computers, implies a recognition that systems quite unlike us superficially might, nevertheless, process information similarly. Hence, many psychological generalizations apply readily enough to them. This fact has two important consequences. First, if parts of our psychological theories usefully describe and explain the behavior and *cognitive lives* of other animals and computers, then the bridge principles necessary to connect some psychological predicates to physical predicates will prove both intractably complex and metaphysically diverse. They must encompass not just disjunctions of a particular human brain's states on different occasions, or even disjunctions of various human brains' and animal brains' states on various occasions. They must also encompass disjunctions of an indefinitely large assortment of machine states (that constitute such cognitive accomplishments as, for example, adding two plus two).

These considerations – conjoined with those arising from the cross-classification of psychological and neural types that results from the diverging agendas of those two sciences – suffice, according to Fodor, to rule out ambitious reductionists' attempts to reduce or dispense with either our psychological theories, the explanations they inform, or the predicates they employ. On this view, psychology is autonomous and unconstrained by neuroscience.

The fit between many of our psychological generalizations and the behavior of computers also has a convenient *strategic* consequence for psychological theorizing. Since, like us, computers can carry out all sorts of cognitive tasks, and since we know just about all there is to know about how computers work, a natural strategy (indeed, some have argued that for computationalists the only plausible strategy) for theorizing about how *we* work is to assume that we work like computers. This assumption introduces the issue of cognitive architecture.

## Cognitive architecture: classicism and beyond

This analogy between humans and computers is cognitive science's preeminent source of theoretical inspiration. Indeed, nearly all theorists hold that at *some* level of abstraction the relation is not mere analogy but identity. This view permits cognitive scientists to explain human cognition by appealing to the concepts and principles of machine computation. Still, beyond a commitment to the notion that cognition involves computations over representations, the precise directions in which this relation should

lead us remain controversial. The emergence of distributed connectionist models over the past decade or so has stimulated debates about the character of both the representations and the computations involved in cognitive processing.

All parties to these debates agree that the nature of the underlying mechanisms restrict the character of the representations and computations in any computational system (though not all agree that the system is computational in the first place). Those mechanisms comprise the *structural* constraints on a (programmable) system's cognitive processing (in contrast to programs' various orchestrations of cognitive functioning). In a computer, these basic mechanisms determine its *functional architecture.*

The concept of cognitive architecture results from applying this notion of the functional architecture of a computer to the human cognitive system. Any computational model of human cognition that aspires to exceed mere input–output equivalence inevitably embodies assumptions about cognitive architecture. The ultimate aim is to specify the constraints which brain mechanisms impose on human cognition. The goal is to provide at least a functional characterization of the basic principles and relations that shape how that neural hardware operates cognitively.

From a computational standpoint, a model of our cognitive architecture should prove *strongly equivalent* with this neural system. The model should not only be input–output equivalent, it should capture the system's primitive representational states *as both primitive and representational,* and it should portray our cognitive processing as transitions between such states (without appealing to other representational states). Although classical and connectionist proposals currently dominate discussions, the space of possible architectures is enormous, leaving room for plenty of new proposals in the future.

Demonstrating a model's empirical accountability requires specifying ways in which human performance can bear on its assessment. For example, architectural features should be cognitively impenetrable. Putative architectural constraints should remain impervious to changes in a person's beliefs; thus, nothing learned should alter architectural constraints. Cognitive scientists have suggested that other types of evidence are relevant as well, including relative sensitivity to damage and chronometric measures of performance.

Unfortunately, additional considerations complicate the evaluation of such evidence. The behavior of a computational system is not just a function of architectural constraints. Programs also play a decisive role. Without extensive knowledge of design, distinguishing those aspects of behavior that arise as a result of the architecture from those that arise as a result of the programs it supports is rarely an easy task – let alone when the systems in question are organic, and the designer is natural selection. When cognitive systems consist of neurons, rather than computer chips, and the designer is evolution, instead of engineers, it is a fairly safe bet that at least sometimes the architecture realizes cognitive functions differently from how digital computers do.

This does not, however, automatically favor distributed connectionist models over classical models of cognitive architecture – certainly not until we know more about the principles guiding neural functioning. Classicism holds that a model of our cognitive architecture provides only a *functional* characterization of the underlying mechanisms. A vast array of physical arrangements can implement the configuration of functional relations which these abstract models describe. On *any* computational view, distinguishing a *cognitive* level from the neuroscientific level of explanation depends

precisely on the fact that models of cognitive architecture involve abstractions away from many of the brain's physical details. Computationalists of both the classical and the connectionist varieties assume that the neural level will *not* prove the best level for characterizing the cognitive architecture. On the classical account, the cognitive level, at which models of cognitive architecture are fashioned, is the lowest analytical level at which states of the system *represent* features of the world. Many connectionists (e.g., Smolensky (1988)) demur – arguably providing more fine-grained analyses of these issues in the process. Brief summaries of classicism and the connectionist alternatives follow.

For the purposes of *theorizing*, proponents of classical models insist on a *principled* subdivision of the cognitive level into a semantic (or knowledge) level and a symbol (or syntactic) level. Theoretical assertions at the semantic level describe human thought in terms of goals and knowledge. As with commonsense psychology, considerations of meaning and rationality order semantic materials. The pivotal assumptions in classical proposals, however, concern the symbol level:

1 Mental symbols are context-independent representational primitives that possess their representational contents by virtue of their *forms*.
2 A finite set of such symbols can represent distinct semantic contents uniquely, because these symbols are the fundamental constituents of a quasi-linguistic system that possesses a concatenative syntax and semantics (that comprehensively parallel one another).
3 The formal, syntactic features of these symbols correspond precisely to neural properties that are pivotal in the etiology of behavior.

The language of thought (LOT) hypothesis (Fodor, 1975) sketches *how* the forms of complex mental representations can coincide point by point with the contents they represent, insuring that no change in content occurs without some change in form. The hypothesis is that they do so roughly as sentences in a language seem to. The forms and the corresponding contents of complex symbolic structures are distinctive combinations of the forms and the corresponding contents of the primitive symbols that are their constituents. The syntactic principles of the brain's computational language are recursive. Because the forms of symbolic expressions uniquely code their representational contents, principles describing the transitions between mental states can be cast syntactically. This is, in effect, to appropriate proof theory from logic to model cognitive processing. Proof theory utilizes a system of syntactic rules for deriving sentences, without appealing to semantics.

Because our mental representations have internal structures, and because principled combinations of primitive mental symbols account for those structures, Fodor and Pylyshyn (1988) insist that thought is:

1 *compositional* – primitive mental symbols are the representational elements from which complex representations are composed;
2 *productive* – although finite in number, they can produce an infinite number of complex mental representations by recursive means; and
3 *systematic* – since, *ex hypothesi*, the forms of the cognitive system's primitive symbols singularly represent their contents, and since the roles they play in the constituent

> structures of complex representations turn completely on those forms, thought is systematic: that is, the ability to entertain some thoughts is intrinsically connected with the ability to entertain others involving the same representational contents.

Finally, classicism holds that the brain states that instantiate the primitive symbols play an essential role in causing our behavior. In addition to the syntactic principles that order them as (representational) primitives, these symbols also submit to neural descriptions that conform to the demands of some eminent, but yet to be imagined, theories in neuroscience. Those theories will identify particular brain states that both instantiate these symbols and exhibit causal relations that match this symbol system's concatenative character. Syntactic principles mediate between our mental states' representational contents and their causal roles, reassuring us that mental representations play just the parts they ought to causally. Thus, the classical account of our cognitive architecture not only provides a framework for preserving our commonsense psychology's explanatory powers; it also envisions a scientific psychology that relies fundamentally on the conceptual framework of that commonsense view.

Connectionist architectures seem to diverge from classical models on nearly every front. They typically consist of a network of simple units in which activation is propagated along connections from input units to one or more layers of hidden units to a set of output units. They do so with neither a program nor a central processor controlling their performance. Frequently, numerous excitatory and (sometimes) inhibitory connections link the units. The links, which are typically feed-forward but can also be feedback (or *recurrent*), have adjustable connection strengths (or *weights*) that influence the amount of excitation or inhibition transferred from one unit to another.

On the basis of excitatory stimuli impinging at the input layer, the units' current levels of activation, and the configuration of all the connections and their weights, connectionist networks produce a pattern of activation (or *activation vector*) at the output layer. Adjusting the connection strengths via feedback learning rules, on the basis of the output vector's divergence from some goal, gradually trains an adequately configured network to respond more appropriately, not only to familiar materials but also to novel materials that manifest similar patterns, thus exhibiting how the system's knowledge resides in its weights. A PDP network's representational capacities, as indicated by its ability to generate appropriate output vectors, regularly involve activity throughout the entire network. Thus, representations are *distributed*. (This is in contrast to *localist* versions of connectionism, which assign semantic contents to specific units.)

These models are frequently introduced by noting their apparent affinities with brain structures, on the assumption that on these fronts, at least, they have an automatic advantage over classical notions of cognitive architecture. Points of similarity between PDP networks and the brain include their parallel processing and distributed representations, as well as their analog capabilities, fault tolerance, and dynamic states. Although no principled barrier precludes either distributed representations or parallel processing in classical architectures, with the exception of the parallel processing in production systems, both are infrequent in classical models.

Perhaps most importantly, connectionist models, unlike classical ones, do not involve the manipulation of symbols according to stored rules. Whatever the preferred interpretations accorded the inputs and outputs of a PDP network, the fundamental

principles that characterize the alterations in individual units' activation levels (which collectively determine networks' trajectories through their state spaces) are mathematical equations that make no appeal to quasi-linguistic forms or semantic contents.

The critical question concerns the relationship between accounts of PDP networks and classical accounts of cognitive processing. Virtually all commentators see important *discontinuities* between the principal analytic categories and explanatory principles that these two accounts employ. Commentators on cognitive architectures, just like commentators on inter-theoretic reduction, part company on the implications of such discontinuities.

Classicists have argued that if connectionist models account for cognitive architecture, then their explanatory principles must appeal to representational states capable of semantic evaluation, which for the classicist also means that they are capable of serving as the constituents of syntactically complex structures. Moreover, if connectionists accept the systematicity of thought, then they must either show that it need not turn on compositionality or demonstrate how connectionist architectures can accommodate that property too.

Fodor and his collaborators maintain that connectionist models don't measure up. Although distributed representations have parts, those parts neither support semantic evaluations nor exhibit the properties of classical constituents. Connectionism lacks the means even to express psychological generalizations that classical theories capture. Thus, they hold that if the explanatory principles of PDP models address brain processes and states, then they do so at an analytical level that is sub-representational and therefore, according to classicism, noncognitive. Instead of characterizing our cognitive architecture, connectionist models only offer information about how a classical architecture might be *implemented* in brains, though even that is only a conjecture.

Fodor and Pylyshyn offer two reasons why matters of implementation are comparatively unimportant. In accordance with the autonomy of psychology, they argue, first, that as a theory of neural implementation merely, connectionism no more constrains cognitive theorizing than do theories from even lower levels, since implementation is a transitive relation all the way down. The *representational* character of mental symbols constitutes a fundamental barrier to the reductionistic program for the theoretical unity of science.

Second, and more importantly, the computer analogy assumes that *cognitive-level* theorizing concerns functional architecture – not the details of its implementations, for the staggering range of physically possible implementations renders them comparatively uninteresting. Fodor and Pylyshyn (1988, p. 63) hold that, in modeling abstract cognitive processes, "there is simply no reason to expect isomorphisms between structure and function" and, more generally, that "the structure of 'higher levels' of a system are [sic] rarely isomorphic, or even similar, to the structure of 'lower levels' of a system." Although, in principle, physical explanations can be had, they supply no insight into *cognitive-level* generalizations; hence, they count only as matters of implementation. The primary charge that connectionism concerns only implementation, then, rests on the same concerns over multiple realizability that generally plague ambitious accounts of reduction.

Advocates of alternative approaches to these issues *acknowledge* the explanatory discontinuities that classicists emphasize; however, they draw quite different conclusions!

Proponents of the dynamical approach diverge most radically from classicism. Employing arguments that, at times, mimic those of eliminativists in the debates about reduction, they envision an account of cognition that *transcends* the notion of representation. They abandon the entire computational project, dispensing, in effect, with the cognitive level and, thereby, with worries about cognitive architecture. Like eliminativists, they are not at all sure that the psychology that classicists defend provides any systematic insights about either the brain or behavior. They focus on networks as dynamical systems, emphasizing the explanatory comprehensiveness and detail of the relevant differential equations. They suspect that the theoretical distance from computation and representation to the mathematics of dynamical systems is unbridgeable. Finally, on this view, even "classic PDP-style connectionism . . . is little more than an ill-fated attempt to find a halfway house between two worldviews" (Van Gelder and Port, 1995, p. 34).

Defenders of connectionism, who put greater stock in our commonsense psychology, still repudiate classicists' defenses of it. Accepting the terms of classicism's challenge, they maintain that even if the eliminativists are right and PDP networks do not instantiate computable functions, they still utilize *representations*. For example, Terrence Horgan and John Tienson (1996) accept LOT and systematicity but deny that either involve a classical, combinatorial syntax. They regard psychological generalizations as *ceteris paribus* laws only and thus as incompatible with the hard rules which classicism requires. Mental representations that classicists regard as *complex* are realized as *primitive* symbols, in the way that irregular past tenses in English (*went*, not *goed*) seem to be (see Article 52, RULES).

While also defending our commonsense framework, Andy Clark (1993) rejects LOT and its accompanying features – certainly as classically comprehended. He too questions classical syntax, noting that PDP researchers have methods for producing structure-sensitive processing without concatenative coding. More generally, if classicism's firm distinctions – for example, between data and processing – do not fit PDP networks, perhaps those distinctions should be *recast*, rather than find connectionism wanting. Considering a flourishing program of connectionist research, let alone a potentially vast collection of yet unanticipated computational devices, why should classical views of basic computational notions, and especially the notion of representation, remain unchallenged? Measures of similarity between representational vehicles in networks can model the semantic similarity of representational contents. The explicitness of a representation need not turn on the tokening of a symbol, but on the ease of use and the multiple exploitability of the information within the system.

Clark emphasizes how representations co-evolve with processing dynamics in human development. That approach renders representational contents dependent upon the processor's capacities and the environment it operates in. Thus, Clark salvages commonsense psychology by attributing to it agendas overwhelmingly inspired by social and cultural practices. For example, he regards the cognitive underpinnings of concepts not as occurrent brain states but as a body of knowledge and skills informing manifestations whose only underlying unity is sociocultural.

Although he stresses the *compatibility* of classical and connectionist insights, Paul Smolensky (1988, 1991, 1995) also contests whether classical conceptions constrain connectionist accounts. He questions the classical assumption that micro-level accounts

of processing details cannot have consequences for conceptions of structural relationships at the level of symbols.

Smolensky notes that connectionist models would implement classical architectures (in the programming language sense) only if classicism provided a precise, comprehensive, algorithmic account of cognitive processing. Anything less means that classical and connectionist models only *approximate* one another. Thus, Smolensky explicitly casts these controversies within the framework of levels of explanation *in science* rather than in programming languages. PDP modeling constitutes a *sub-symbolic paradigm* operating at the *sub-conceptual level*, which falls between the conceptual and the neural and currently offers the best means for theoretically connecting symbolic computation to neural functioning.

Smolensky maintains that such inter-level interaction can lead to the improvement of higher-level theories, hence sub-symbolic research can refine the classical approach. Employing processing algorithms affording greater precision and detail, connectionists can relate activity patterns to conceptual-level descriptions. Such integrative research across explanatory levels results in the *reconceptualization* of the notion of cognitive architecture. Smolensky advocates an "intrinsically split-level cognitive architecture" (1991, p. 204). *Syntactic* relations are characterized in terms of algorithms that describe the alterations in individual processing units' activation levels, while semantic interpretation transpires in terms of larger activity patterns. Sub-symbolic analyses offer new formal instantiations of computational concepts. Smolensky emphasizes that explicating classical notions in the language of continuous computation relies on a semantic shift accompanying the shift to the sub-conceptual level.

Smolensky's claim that sub-symbols, as activity patterns in networks, correspond to symbolic constituents has stirred debate. His critics insist that this constituency is not classical. Smolensky responds that sub-symbolic accounts provide penetrating *approximations* of compositional structure and LOT. The pivotal question is how the variable activities in networks achieve symbols' representational stability. Smolensky replies by turning this problem on its head, noting how connectionist nets readily accommodate the *context sensitivity* of representations (for which considerable psychological evidence exists). Sufficient representational stability depends not on symbolic form but on a *family resemblance* (1988, p. 17) among those vectors that, in different contexts, carry out some functional, sub-symbolic role.

Smolensky (1995) has elaborated an integrated connectionist/symbolic (ICS) architecture with which he aims, ultimately, to surmount any simple distinction between classical and connectionist architectures. Smolensky's ICS architecture employs general PDP principles constrained by tensor product structures that insure that both the semantics and the functions to be computed can be managed symbolically, even though symbols play no causal role in the computations. Harmonic nets, which are structured to maximize parallel soft-constraint satisfaction or *harmony* gradually, realize the various higher cognitive processes that symbolic accounts describe. Smolensky argues (1995) that such a sub-symbolic reduction motivates revisions in symbolic accounts that enable a richer theoretical integration of the two levels, resulting in the *preservation* of classical insights. Smolensky offers harmonic grammars and optimality theory's contributions to syntactic studies and phonology as illustrations of revisions that enrich classical accounts and preserve their most important claims.

Smolensky anticipates the same sort of approximate reduction arising from the co-evolution of theories at different levels that various philosophers have championed. Co-evolving theories often yield progressively better inter-theoretic mappings. The increased integration associated with such co-evolution will not eliminate or replace symbolic accounts, but rather improve research at the conceptual level.

Smolensky's comments that any definition of constituency that provides "explanatory leverage is . . . valid" and that classical architecture is a "scientifically important" approximation of the underlying dynamics at the sub-conceptual level (1991, pp. 210, 203) accord with the *pragmatism* of recent *integrative models* of cross-scientific relations. Increasingly, philosophers argue that the welcome simplicity associated with reductionism exacts too high a price. Reductionism neglects all relations between explanatory levels except those between theories, and it conceives all inter-theoretic relations in terms of reductive explanation. Compared to reductionist accounts, integrative models explore a wider range of just the sort of cross-scientific relations that are particularly prominent in interdisciplinary research typical in cognitive science. Examining issues of discovery, evidence, method, and more, advocates of integrative models foresee *many* illuminating relationships (besides possible reductions) between psychological, connectionist, and neuroscientific models.

William Bechtel and Robert Richardson (1993) argue that the chief goal of reductionistic research among practicing scientists is the discovery and explication of the *mechanisms* underlying the functioning of complex systems. Pursuing the strategies of structural decomposition and functional localization, scientists steadily unveil the various micro-level mechanisms realizing higher-level patterns. This activity neither eliminates nor replaces the complex system or macro-level theories.

Smolensky notes that considerations of mathematical modeling, more than neural considerations, drive developments in connectionist research. He also emphasizes the accuracy, precision, and comprehensiveness of dynamical systems theory as an account of connectionist processing. He nonetheless conceives of connectionist modeling as a kind of primordial *neuro*computational research. *Contra* Van Gelder and Port (1995), developing theories of a system's dynamical features at one analytical level does not usually warrant ignoring theories of that system's parts and structures at that or higher levels (Clark, 1997). Explanatory levels contain theories of a system's synchronic and diachronic dimensions. Integrative models propose that interactions between research on synchronic and diachronic matters at a single level and between research of either sort at different levels are *mutually* enriching.

Research at lower levels can refine and even correct higher-level approximations. But integrative models also show how upper-level research (e.g., in psychology) can play a significant role in *justifying* lower-level proposals and motivating innovative research at intermediate levels. Attention to the psychological evidence *enhances* the precision and plausibility of connectionist and neuroscientific models (McCauley, 1996).

Valerie Hardcastle (1996) stresses how these interdisciplinary endeavors stimulate research in *bridge sciences* (such as evoked-response potential studies) and contribute to the *explanatory extension* of the sciences involved, either by conceptual refinement or by the theoretical support of one science for an antecedently problematic assumption of an other. Hardcastle criticizes classicists' assumptions about the clarity of

distinctions between structures and functions that are so pivotal to their sharp distinction between architecture and implementation. She argues that whether a description counts as structural or functional depends upon the analytical levels involved, the questions asked, the related explanations available, and the background knowledge at hand. What might look like implementational detail from a higher-level perspective (e.g., different measures of clustering in network activity yielding different accounts of conceptually interpretable patterns) may have architectural implications from a lower-level perspective (if, for example, the differences among these measures' accounts of such patterns are found to turn systematically on micro-level variables).

Multiple realizability does not necessarily present intractable problems for integrative models. Alternative realizations of psychological states raise neither barriers to cross-scientific connections nor grounds for declaring disciplinary autonomy, but opportunities for further empirical research about the complexity of the interface between the psychological and the neural. If something like the identity theory were to prove plausible even for some extremely limited cognitive domain, the possibility of alternative realizations will certainly not deter scientists from exploring and exploiting all the resulting cross-scientific connections! If multiple instantiation of psychological functions proves the rule, it does not follow that – and in many cases there is little reason to expect that – neuroscientists face an unmanageably large number of alternatives. *Even* if token physicalism is basically correct, the important question for integrative models is whether it might sustain some cross-scientific connections that advance research in cognitive science. Unlike most reductionists and many of their prominent critics, integrative modelers do not presume that the answer to that question can be determined on principled grounds.

## References and recommended reading


*Bechtel, W. (ed.) 1986: *Integrating Scientific Disciplines*. The Hague: Martinus Nijhoff.

Bechtel, W. and Richardson, R. C. 1993: *Discovering Complexity*. Princeton: Princeton University Press.

*Bickle, J. 1995: Connectionism, reduction, and multiple realizability. *Behavior and Philosophy*, 23, 29–39.

Churchland, P. M. 1989: Eliminative materialism and the propositional attitudes (1981). Repr. in *A Neurocomputational Perspective*, Cambridge, Mass.: MIT Press, 1–22.

Clark, A. 1993: *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge, Mass.: MIT Press.

—— 1997: *Being There: Putting Brain, Body, and World Together Again*. Cambridge, Mass.: MIT Press.

Fodor, J. A. 1975: *The Language of Thought*. New York: Thomas Y. Crowell Co.

Fodor, J. A. and Pylyshyn, Z. W. 1988: Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28, 3–71.

Hardcastle, V. G. 1996: *How to Build a Theory in Cognitive Science*. Albany, NY: SUNY Press.

Horgan, T. and Tienson, J. 1996: *Connectionism and the Philosophy of Psychology*. Cambridge, Mass.: MIT Press.

McCauley, R. N. 1986: Intertheoretic relations and the future of psychology. *Philosophy of Science*, 53, 179–99.

—— 1996: Explanatory pluralism and the co-evolution of theories in science. In R. N. McCauley (ed.), *The Churchlands and their Critics*, Oxford: Blackwell, 17–47.

*McLaughlin, B. P. 1997: Classical constituents in Smolensky's ics architecture. In M. L. Dalla Chiara et al. (eds), *Structures and Norms in Science*, Dordrecht: Kluwer Academic Publishers, 331–43.

Smolensky, P. 1988: On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11, 1–74.

—— 1991: Connectionism, constituency, and the language of thought. In B. Loewer and G. Rey (eds), *Meaning in Mind: Fodor and his Critics*, Cambridge, Mass.: Blackwell, 201–27.

—— 1995: Reply: constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In C. MacDonald and G. MacDonald (eds), *The Philosophy of Psychology: Debates on Psychological Explanation*, Oxford: Blackwell, 223–90.

*Van Gelder, T. 1995: What might cognition be if not computation? *Journal of Philosophy*, 91, 345–81.

Van Gelder, T. and Port, R. F. 1995: It's about time: an overview of the dynamical approach to cognition. In R. F. Port and T. van Gelder (eds), *Mind as Motion*, Cambridge, Mass.: MIT Press, 1–43.