Entry on "Philosophical Issues in Brain Theory" for the
HANDBOOK OF BRAIN THEORY AND NEURAL NETWORKS (ed, M. Arbib)
MIT Press, 1995 738-741
(Fully revised and expanded entry, now co-authored with
Chris Eliasmith, to appear in the second edition, June 2000)

# Philosophical Issues in Brain Theory and Connectionism (Revised Version, 2002)

*Andy Clark & Chris Eliasmith*

## Introduction

In this article, we highlight three questions: (1) Does human cognition rely on structured internal representations? (2) How should theories, models and data relate? (3) In what ways might embodiment, action and dynamics matter for understanding the mind and the brain?

The first question concerns a fundamental assumption of most researchers who theorize about the brain. Do neural systems exploit classical compositional and systematic representations, distributed representations, or no representations at all? The question is not easily answered. Connectionism, for example, has been criticised for both holding and challenging representational views. The second quesútion concerns the crucial methodological issue of how results emerging from the various brain sciences can help to constrain cognitive scientific models. Finally, the third question focuses attention on a major challenge to contemporary cognitive science: the challenge of understanding the mind as a controller of embodied and environmentally embedded action.

## Does cognition need representations?

"Does cognition need representations" is the most difficult and least well defined of our three questions. But it addresses one of the most philosophically interesting features of many connectionist models, especially those most closely related to brain theory. The intuúition is that connectionism poses a challenge to the classical view of the brain as a syntax-sensitive engine. This idea involves depicting all or most of human cognition as involving something akin to logical operations apúplied to something akin to linguistic (sentential) structures. The prime philosophical exponent of the identification of genuine cognitive processes with such operations on quasi-sentential entities is Fodor (see Fodor, 1987; Fodor and Pylyshyn, 1988). Fodor argues in favor of an innate symbolic code (the *Lanúguage of Thought)* and of mental processes as involving operaútions defined over the syntactically structured strings of that code. The underlying image is of an inner economy in which symbol strings are operated on by procedures sensitive to the structure of the string.

In contrast, a typical trained-up network, does not employ grammatical strings nor, *a fortiori*, processing operations sensitive to the structure of such strings. Instead, we find prototypical complexes of properties represented in a high-dimensional space (see P.M. Churchland 1995). The space is highly organized in the sense that data items which need to be treated in closely related ways become encoded in neighboring regions of the space. It is this *semantic metric* which allows the network to generalize and to respond well under conditions of noise, etc. But this organization of the encoded knowledge does not amount to the provision of a genuine syntax. One way to see this difference is to ask what rules of combination of represented elements apúply, and in what systematic ways we can operate on complexes of represented items. As Fodor and Pylyshyn (1988) point out, there is no analog

in distributed representations to the logical operations of detaching an element from one string (complex representation) and adding it to another.

Nevertheless, a variety of connectionist techniques have been developed to allow for structure-sensitive processing, but such techniques have been deúscribed (van Gelder, 1990) as providing *functional,* as opposed to *concatenative,* compositional structure. A complex represenútation has concatenative structure if it embeds the individual constitutive elements unaltered within it. It has functional comúpositional structure if such components are usable or retrievúable, but the complex expression does not itself embed unalútered tokens of these parts. Most connectionist schemes for dealing with compositional structure are functionally compositional (e.g., RAAM architectures, tensor product enúcodings, holographic reduced representations (HRRs); although synchrony binding is concatenative. For a review, see CONNECTIONIST AND SYMBOLIC REPRESENTATIONS). Of these, HRRs are perhaps best suited to bridging the traditional gap between connectionist and symbolicist approaches to understanding language-like processing. HRRs are supremely structure sensitive, do not suffer from the dimensional increases of tensor products, and can be implemented in standard connectionist networks, yet they are not concatenative (Eliasmith and Thagard, 2001).

In our opinion, a major benefit of exploring the space of connectionist cogúnitive models is thus that it may help us expand our sense of the possible nature of internal representation and hence better understand what is truly essential to notions such as *structure, synútax, and complex representation.* Doing so, we may discover which aspects of our models are simply artifacts of our (over)familúiarity with one representational format, viz., the format of atomic elements and grammar common to language and logic.

## How do theories, models and data relate?

As a computational formalism connectionism is quite powerful, allowing us to approximate nearly any function or performance profile that we desire. However, the mere fact that some input-output pattern *P* is found in human cognition and can be mimicked using some connectionist model is, in itself, of only marginal psychological interest. The demands of cognitive science, unlike those of artificial intelligence, require more. Ideally, we must provide models which are both consistent with neurological data and comprehensible. However, the relation between data and models is not uni-directional. Models are constrained by data, but they also help us determine what sorts of experiments to use in looking for more relevant details. The role of theories in unifying the brain sciences and connectionism is an important, albeit (for the time being) a mysterious, one (although see Eliasmith and Anderson, in press, chp. 1 for one possibility).

The data which constrain models comes largely from two sources: higher levels, from work in disciplines such as psychology and psycholinguistics, and lower levels, from work in neuroscience and brain theory. From psychology and psycholinguistics we can extract vast bodies of constraining data which go way beyond the mere specification of a task-specific input-output mapping. Such data can concern, for example, the relative difúficulty of parsing certain sentences or solving certain problems, the time course of problem solving, the developmental profile of skill acquisition, and the way in which new and old knowlúedge interacts in the context of new learning (for detailed exúamples, see Karmiloff-Smith, 1992; DEVELOPMENTAL DISORDERS).

For current purposes, however, it is the lower level constraints which we seek to highlight. The question here concerns the

propúer relation between connectionist computational modeling and the detailed constraints emerging from the various brain sciúences. Such sciences include neuroanatomy, neurochemistry, lesion studies, and research on the single cell, circuit, and sysútems level. It seems clear that any acceptable model of human information processing must respect the results of such studies. To do so, some intelligible relation must exist between the theories put forward by, for example, connectionist computaútional modeling and the entities and lawful interactions studied by the brain sciences. It is a duty sadly neglected by both classical artificial intelligence and a great deal of connectionist work to make some effort to display the precise nature of such relations.

Such a task is complicated by the variety of levels of interest which may characterize the brain sciences. These include the levels of biochemical specification: single cells, circuits, subsysútems, and networks of subsystems. Marr's suggestion that studies at each level can be independently pursued is highly dubious. Our top-level decomposition of a task into subtasks apt for computational modeling may be challenged once we become familiar with the distribution of information processing resources in the brain. What we originally thought of as two distinct functions may actually share circuitry in the brain (see Arbib, 1998). Such a result will not be devoid of psychological significance, since it will figure in an explanation of the breakdown profile as reúvealed by, for example, lesion studies of the system.

How, then, should we conceive the bridge between idealized artificial intelligence models and brain theory? It is precisely the complex relations between implementation and function that have spawned a recent surge of interest in computational neuroscience. With the explicit goal of taking biological constraints as seriously as computational ones, computational neuroscience has begun to explore a vast range of realistic neural models. For example, Reike et al. (1997) provide an information theoretic analysis of

spike trains, allowing accurate stimulus signal reconstruction. The combination of such spike train analyses and, for example, Abbott's (1994) higher-level discussions of basis function representations, can provide valuable insights into the functioning of populations of neurons (see Eliasmith and Anderson, in press). Though preliminary, the tools developed by such research are promising candidates for generating biologically realistic connectionist models.

Such models should prove useful in providing constraints of their own. Insights from basis function analyses suggest new experiments for neurophysiologists. In particular, it seems that neurons may have higher-dimensional tuning profiles than previously imagined. Though neurological techniques for determining complex profiles have yet to be perfected, connectionist modeling suggests such tuning properties are important to the everyday functioning of neurons. So, not only does biology inform the construction of computational models but, ideally, those same models can help suggest important experiments for neuroscientists to perform. In this sense, models and data can be mutually beneficial. Of course, the benefits are highly constrained by assumptions of both the model and experimental design.

Although no model can be expected to do justice to all aspects of its target, what justice it can do depends on the biological realism of the assumptions behind the model. Biological realism, of course, can be incorporated into a model in many ways; e.g., by including neurochemical diffusion, single neuron morphology, spike train statistics, neuroanatomical constriants, population dynamics, or system-level organization (see Eliasmith and Anderson, in press, for examples). In any case, what we can and should expect from a modeler is a clear statement of what aspects of the target phenomenon are supposed to be explained, and (if it is a computational model) at what level the computational story is

intended to capture real neurophysiological facts. Successful attempts to exploit the close relation between experiment and model are still something of a rareity. This is largely because theoreticians (typically mathematicians, physicists and engineers) and experimentalists (typically neuroscientists and biologists) do not yet have many conceptual tools in common. In order to reap the benefits of mutual, inter-level constraint, this will likely have to be rectified.

## In what ways might embodiment matter for understanding the mind and the brain?

In recent years, an important challenge has been issued to cognitive science. It stems from the work of researchers espousing the *dynamicist hypothesis* (van Gelder, 1995). The dynamicist commitment to making time central to cognitive modeling is inspired by the broader realization that cognitive systems are real physical systems acting in the real world in real time. Given the finite, though vast, computational resources of the brain, it also seems that evolution has often off-loaded complex computational tasks to the body and to the environment. This double 'situatedness' of cognitive systems needs to be reckoned with if we are to develop an accurate picture of precisely the kinds of computation neural systems perform. Connectionism and brain theory must conspire to explain this kind of representational and computational economy. Thus, while looking *inside* to the brain and the results of neuroscience we can not afford to turn a blind eye to constraints and resources which come from the *outside*, the gross body and environment of a cognitive system (Clark 1997).

Consider vision. There is now a growing body of work devoted to so-called 'Animate' vision (Ballard 1991). The key insight here is that the task of vision is *not* to build rich inner models of a surrounding 3D reality, but rather to use visual information

efficiently and cheaply in the service of real-world, real-time action. Animate vision thus rejects Marr's analysis, what Churchland et al. nicely dub the paradigm of "pure vision" – the idea (associated with work in classical AI and in the use of vision for planning), that vision is largely a means of creating a world model rich enough to let us "throw the world away', targeting reason and thought upon the inner model instead. Real-world action, in these 'pure vision' paradigms, functions merely as a means of implementing solutions arrived at by pure cognition.

The Animate vision paradigm, by contrast, gives action a starring role. Computational economy and temporal efficiency is purchased by a variety of bodily action and local environment exploiting tricks and ploys including

〈 the use of cheap, easy-to-detect (possibly idiosyncratic) environmental cues (e.g., Searching for Kodak film in a drug store? Seek 'Kodak yellow'.);

〈 the use of active sensing (e.g., use motor action, guided by rough perceptual analysis, to seek further inputs yielding *better* perceptual data – move head and eyes for better depth perception, etc.); and

〈 the use of repeated consultations of the world in place of rich, detailed inner models.

Ballard et al. (1997) have recently demonstrated that subjects do not bind color and location information in a block-copying task until it is absolutely required by current problem-solving. As a result, changes made to the display (such as switching the color of blocks during a saccade) are very often undetected.

Vision, this body of work suggests, is a highly active and intelligent process. It is not the passive creation of a rich inner model, so much as the active retrieval (typically by moving the high resolution fovea in a saccade) of useful information *as it is needed* from the constantly present real-world scene. Ballard et al. speak of "just-in-time representation", while the roboticist, Rodney Brooks, has coined the slogan "The world is its own best model" (Brooks 1991). The combined moral is clear: vision makes the most of the persisting external scene, and gears its computational activity closely and sparingly to the task at hand.

The general thrust of the Animate vision research program, however, is not to reject the ideas of internal models and representations, so much as to reconfigure them in a sparser and more interactive image. We thus read of inner databases that associate objects (e.g., my car keys) and locations (on the kitchen table), of internal feature representations, of indexical representations and so on. What is being rejected is not the notion of inner content-bearing states per se, but only the much stronger notion of rich, memory-intensive, all-purpose forms of internal representation.

The crucial distinction, it seems to us, is thus not between representational and non-representational solutions so much as between rich and action-neutral forms of internal representation (which may increase flexibility but require additional computational work to specify a behavioral response) and sparse and action-oriented forms (which exploit the body and world and which begin to build the response into the representation itself).

## Discussion

Our vision of basic biological reason is rapidly changing. There is a growing emphasis on the computational economies afforded by

real-world action and our growing appreciation of the way larger structures (of agent and artifacts) both scaffold and transform the shape of individual reason. These twin forces converge on a rather more minimalist account of individual cognitive processing – an account that tends to eschew rich, all-purpose internal models and sentential forms of internal representations. Such minimalism, however, has its limits. Despite some ambitious arguments, there is currently no reason to doubt the guiding vision of individual agents as loci of internal representations and users of a variety of inner models. Rather than opposing representationalism against interactive dynamics, we should be embracing a broader vision of the inner representational resources themselves.

The sciences of the mind are thus in a state of productive flux: the product of multiple converging influences coming from real-world robotics, systems-level neuroscience, cognitive psychology, evolutionary theory, Artificial Intelligence, and philosophical analysis. This flux has forced us to reconsider earlier accounts of the relation between thoery, models, and data relevant to cognitive systems. More importantly, we can see a new vision of mind emerging. The point at which many of these influences currently converge is captured by seeing mind as *in essence* a controller of embodied and environmentally-embedded action. Mind is an organ for orchestrating real-time responses to a real world.

One major player in these recent events has been the explosion of work on Artificial Neural Networks. Such networks amounted to an existence proof of the possibility of adaptive intelligent behavior without reliance on explicitly formulated rules or language-like data-structures. Moreover, the networks integrated representation and action in a very direct manner: knowledge became encoded in a form dictated by its use in a particular type of problem solving. But the Neural Networks revolution was incomplete. It was incomplete because it was still burdened with much of the unnecessary baggage of the previous, disembodied,

symbol-crunching approach to understanding cognition. Mind was still treated as an essentially timeless locus of abstract problem-solving capacities.

All this changed with the surge of interest (in the late 80's, early 90's) in what became known as Autonomous Agent research (see e.g., essays in Beer, Ritzmann and McKenna (1993)). This research aimed to model and understand the adaptive success of single, complete, embodied systems: insects which walk and seek food, the cockroach's amazingly sophisticated mechanisms for detecting and evading attackers, robots which learn to swing from branch to branch using real mechanical arms, etc., etc.. Many of these models exploit Artificial Neural Networks as control systems. But the constraints on success became very different.

Finally, the constraints on computation using Artificial Neural Networks are very different from those on real biological computation. It is here that the relation between theory, model, and data again becomes pivotal. Interestingly, reconceptualizing mind in each of these previous cases has depended on rethinking the relevant constraints (i.e., linguistic vs. non-linguistic symbols, partial vs. full-bodied systems). Introducing the complexities of natural neural computation is bound to have a similar impact on our concept of mind.

So, many important questions remain. Can work in artificial neural networks come to grips with the real complexity of biological computation? What kinds of systems-level model can help make sense of the complex balance between specialization and co-operation that we find in real brains? Can a representation-sparse approach make headway with all aspects of human cognition, or is it limited to cases of perceptuo-motor control and on-line reasoning? How does the command of public language impact and transform human thought and reason?

The cognitive science of the biological, embodied mind is still in its infancy, and the full power and scope of the new vision remain to be determined. But the issues raised will, we believe, shape the agenda of the next decade of research into mind and its place in nature.

**Road Map**: Connectionist Psychology

**Related Reading:** Artificial Intelligence and Neural Networks; Conúsciousness, Theories of; Perspective on Neuron Model Complexity; Structured Connectionist Models

## References


Ballard, D. (1991) Animate Vision Artificial Intelligence 48, 57-86

Ballard, D., Hayhoe, M., Pook, P. and Rao, R. (1997) "Dieictic Codes For the Embodiment of Cognition" Behavioral and Brain Sciences 20:723-767

*Beer, R., Ritzmann, R., and McKenna, T. (Eds.) (1993) Biological Neural Networks In Invertebrate Neuroethology and Robotics. London: Academic Press.

Brooks, R. (1991) Intelligence without representation Artificial Intelligence 47, 139-159

*Churchland, P.M. (1995) The Engine of Reason, The Seat of the Soul (MIT Press, Camb. MA)

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

*Clark, A., 1997, Being there: Putting brain, body and world together again. Cambridge, MA: MIT Press.

Eliasmith, C. and P. Thagard, 2001. Integrating structure and meaning: A distributed model of analogical mapping. Cognitive Science, 25(2), 245-286.

*Eliasmith, C. and C. H. Anderson, in press, Simulating Neurobiological Systems: Principles and Methods, Cambridge, MA: MIT Press.

*Fodor, J., 1987, Psychosemantics: The Problem of Meaning in the Philosophy of Mind, Cambridge, MA: MIT Press.

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

Karmiloff-Smith, A., 1992, Beyond Modularity: A Developmental Perúspective on Cognitive Science, Cambridge, MA: MIT Press.

Reike, F., Warland, D., de Ruyter van Steveninck, R. and Bialek, W., 1997, Spikes: exploring the neural code. Cambridge, MA, MIT Press.

van Gelder, T., 1990, Compositionality: A connectionist variation on a classical theme, Cognitive Science, 14:355-384.

van Gelder, T., 1995, What might cognition be, if not computation? The Journal of Philosophy 91(7): 345-381.