Minimal Rationalism

Andy Clark
School of Cognitive and Computing Sciences
University of Sussex
BRIGHTON BN1 9QH
andycl@cogs.susx.ac.uk

Abstract

Enquiries into the possible nature and scope of innate knowledge never proceed in an empirical vacuum. Instead, such conjectures are informed by a theory (perhaps only tacitly endorsed) concerning probable representational form. Classical approaches to the nativism debate often assumes a quasi-linguistic form of knowledge representation and delineate a space of options (concerning the nature and extent of innate knowledge) accordingly. Recent connectionist theorizing posits a different kind of representational form, and thus determines a different picture of the space of possible nativisms. The present paper displays this space and focuses on an especially interesting sub-region labelled "Minimal Rationalism". The philosophical significance of the minimal rationalist option is explored. Two consequences which emerge are first, that the apparently clear distinction between innately specified knowledge and innately specified structure is shown to be unproductive; and second, that there may exist tracts of innate knowledge whose content is not propositionally specifiable.

0. Nativism. Why worry?

Sometimes trivial, usually fruitless, the Nativism/non-Nativism debate generally ends not with a conclusion but with a whimper. All parties agree that something important is present in us without being the product of genuine individual learning. All that then remains is to determine what. And that, as has been vigorously argued in the past (e.d. Fodor (1980)) is in the end an empirical question whose detailed answer is not to be determined by armchair philosophical speculation. Most of the published debate thus consists in arguing about whether some of our innate endowment is highly domain-specific (e.g. Chomsky (1986)) or instead relates to basic, general-purpose problem solving (e.g. Putnam (1981)). A second major strand of the published debate relates specifically to concepts and revolves around the question whether anything genuinely worth calling concept learning actually takes place, or whether all our conceptual repertoire must be in some non-trivial sense innate (Fodor (1980) and papers in Piatelli-Palmarini (1980)).

The present treatment maintains a safe distance from these types of question (a few asides notwithstanding). Instead, the focus is on the way in which the possibility of innate knowledge is conceived. I shall argue that the received conception of the space of possible options is in fact a product of the (often tacit) acceptance of a certain model of the probable form of internal knowledge representation: a form whose clearest expression is found in the hypothesis of an innate language-like representational system (a Language of Thought). Change the conception of the form of internal presentation and you radically alter (or so I shall argue) the picture of the space of possible options.

This potential alteration has not gone unnoticed in the recent literature. Important treatments include Ramsey and Stich (1991), Narayanan (1992) and Karmiloff-Smith (1992a). Several of the themes I develop in sections 1-3 - where I discuss the impact of connectionism on the nativism debate, in the broadest terms, are rooted in these exploratory forays. The remainder of the paper, however, tries to push the new debate a little further. Thus section 4 introduces (with some simulation results) a largely unnoticed (but see Karmiloff- Smith (1992a)(1992b)) yet potentially highly significant possibility which I term 'Minimal Rationalism'. A minimally rationalist innate endowment involves the (domain-specific) pre-setting of tiny but vital information-processing parameters which, in a delicate co-operation with predictable environmental inputs, result in the acquisition of specific items of knowledge. To understand the nature of such minimal endowments we need to use a new set of tools. Instead of conceptualizing any genuine innate knowledge as consisting in familiar kinds of conceptual or propositional content, we need to move towards a more 'geometric' understanding. In particular, we need to exploit the idea of an error surface determined by the setting of numerical parameters in a high-dimensional space. The specification of innate knowledge, I shall argue, will often consist (necessarily!) in the fixation of a favourable position on such an error surface. Once we thus expand our notion of innate information beyond the realms of what is in-principle propositionally specifiable, it becomes increasingly difficult (section 5) to separate questions concerning the innate structure (e.g. the local architecture (of layers, modules etc.)) of a computational subsystem from questions concerning innate knowledge. Classical treatments of the nativism debate could support such a separation since they allowed a sharp distinction between computational profile (algorithm and data) and implementation (the underlying physical device). Connectionist approaches erode that distinction and hence blunt the difference between structure, algorithm and information.

1. Nativism and Representational Form

It is no accident that much of the historical debate concerning the pros and cons of nativism revolved around the notion of an innate idea. For talk of ideas, vague thought was (and is) nonetheless reflected the best available theory of that in which our mature knowledge might consist. And our conception of the potential nature of any innate endowment was, by default, modelled on our conception of the nature of the mature product.

In talking of innate ideas in the mind, we are not yet forced to consider questions concerning any possible physical vehicles for those ideas. In these more rampantly physicalist times, however, questions concerning the possible contents of tracts of innate knowledge have been inspired not just by a vision of the contents of the mature product but also by a vision of the form of their inner vehicles. The clearest example of this line of influence is seen in the works of Jerry Fodor.

Fodor subscribes to what I shall call 'Bipartite Nativism'. Such a nativism ascribes two types of innate endowment to the human neonates. These are:

1. An innate (but peripheral) system of processing modules which are significantly structured so as to promote the acquisition of specific skills (e.g. grammar acquisition).
   (see Fodor (1983)).

2. An innate (and central) corpus of representational atoms (which includes atomic items corresponding to most lexical concepts and which merely require triggering by exposure to appropriate environmental stimuli).
   (see Fodor (1975), (1980), (1987)).

Fodor thus subscribes to both a kind of 'gross architectural' nativism (for the modules) and a 'symbolic nativism' (for central processing).

In the following sections I shall try to articulate a very different picture. It is a picture in which the image of the form of representation of mature knowledge (of the kind which Fodor would ascribe to 'central processing') is very different. This difference, I shall argue, leads us to reconceive the notion of innate knowledge in important ways and eventually blurs the architecture/representation distinction itself.

2. Connectionism: The Bare Essentials

The broad lines of the Connectionist Cognitive Paradigm are by now familiar to most philosophers (for introductory treatments see Clark (1989), Bechtel and Abrahamson (1991) and the essays collected in McClelland, Rumelhart and the PDP Research Group (1988) vols. I and II)) and I shall risk only a summary introduction here. It is in any case the specific vision of the form of any innate endowment which is going to do most of the work in what follows. Still, a few words are in order.

The connectionist approach, insofar as it presents itself as a genuine alternative to classical 'rule and symbol' systems, relies on (i) an alternative form of knowledge representation, (ii) an alternative type of basic processing operation and (iii) a set of powerful learning algorithms.

Regarding knowledge representation, the radical connectionist eschews representations which consist of symbolic atoms concatenatively combined to form symbolic expressions. (For a good discussion, see Van Gelder (1990)). Instead, connectionism exploits activation patterns among large numbers of idealised 'neurons' (small processing units) to encode specific contents. The resulting scheme turns out to resemble prototype based encoding insofar as similar contents tend to be represented by similar patterns of activation (hence the inner 'symbols' are in a sense non-arbitrary: if content A is represented as pattern of activation P, it will be semantically significant if a content B is assigned a closely related pattern (see Clark (forthcoming a) Chapter 2 for a full discussion). All the semantically significant items in such an encoding can thus have significant internal structure. In a very real sense, there are no symbolic atoms here i.e. no items which are both clearly representational and lack semantically significant inner structure. Moreover, complex contents are not represented by concatenations of more basic representations but by new activation patterns (ones which need not literally embed the 'components') created by processes involving mathematical operations upon the numerical vectors which constitute the aforementioned 'activation patterns'. Once again, the departure from the classical paradigm is quite marked (see Smolensky (1909), Fodor and McLaughlin (1991)).

In such systems, the basic processing operations are defined over such numerical vectors. Information retrieval consists in a process of vector completion given a partial vector as a cue. Generalization is achieved by the superpositional storage of activation patterns in a single set of long term weights. The weights consist of numerical values assigned to local links between idealised neurons. It is these weights which allow the system, given a partial vector (pattern of activation across a set of input units) as a cue, to complete the vector (by activating, courtesy of the connection weights, a specific pattern of units). If several contents are stored superpositionally in a single network of units and weights, an input cue which is appropriate to several such patterns will induce an activation pattern which in a sense averages the patterns of the individual contents which fit the cue. Hence so-called 'free generalization' (see Churchland (1989) ch.9).

Finally, and most significantly, such networks are heir to some powerful learning algorithms. Starting with random weights on the connections a network can automatically alter these random weights in a way which should lead it to encode a desired input-output mapping. This kind of learning is usually driven by exposing the net to a set of inputs alongside a set of desired outputs. The net uses the (initially random) weights to yield an (initially hopeless) output. If the output is incorrect, an automatic procedure slightly amends those weights most heavily implicated (along the path of activation between input and output) in the mistake in whatever direction (increase or decrease specific weights) will yield a reduction in the numerical error measure. Such a process (of 'gradient descent learning' - see e.g.

P.S.Churchland and T.Sejnowski (1992) pp.106-7) gently leads the network in the direction of an assignment of weights which will support the target input-output mapping and (usually) will generalize to deal with new cases of the same type (e.g. a net trained to map coding for written text to coding for phonemes will then perform the mapping for text on which it was not specifically trained - see Sejnowski and Rosenberg 1986) (1987)).

Even such a summary sketch succeeds (I hope) in displaying the genuine distance which separates these connectionist models from their classical cousins. Where classicists were tempted (maybe even forced - see Fodor (1975)) to posit a system of innate symbolic atoms and significant innate architectural structures (the modules of Fodor (1982)) the connectionist may appear ready to reject both: to insist on a single network of units and weights and to begin with random weights and hence no ready-made set of symbolic atoms. But this, as other commentators have rightly pointed out (see e.g. Churchland (1989), Karmiloff- Smith (1992a), Narayanan (1992) would be way too hasty. The connectionist (like everyone else from behaviourists upwards - see e.g. Quine (1969) p.96) must often be a nativist too. But the empirical details of the connectionist approach determine a space of nativist options which is importantly different to the classical space. I shall sketch that space, and then proceed to a thorough investigation of my favoured corner of it; a subspace I term 'minimal rationalism'.

3. The Space of Connectionist Nativisms.

The space of possible connectionist nativisms is bounded by two extremes. One extreme is the Connectionist Tabula Rasa: a single, big undifferentiated network which begins with a random assignment of weights. The other extreme is the Connectionist Classical Device: a units-and-weights style implementation of the full bipartite classical story, with innately specified modules and a central system which uses connectionist resources to implement a full classical symbol system. (For a sketch, see Touretsky and Hinton (1985), Touretsky (1989).) The Connectionist Classical device we put aside. It is of little philosophical interest in the present context. The Connectionist Tabula Rasa, although it is shortly to be rejected (on empirical grounds) merits a few initial comments.

First, and most obviously, the connectionist Tabula rasa (like its associationist ancestors) is not a totally blank system after all. For it comes equipped with both a structure (a specific number of units and weights, and a specific configuration into input layers, output layers and intervening layers) and a learning rule. This is unsurprising. As Samet (1986) comments "Even tabulas have some innate structure" (p.575). The Connectionist Tabula Rasa is not, anyway, to be taken seriously as a model of the human neonate's cognitive state. A wealth of results in psychology and neuroscience attests to the significant amounts of additional innate structure upon which human cognition relies (see e.g. Churchland and Sejnowski (1992)). And working connectionists have come to appreciate more and more the need to pre-structure networks to perform complex tasks -see e.g. Plunkett and Sinha (1992), McClelland, J.L. (1989), Le Cun et al. (1989). All that said, there is still an important existence proof embodied in the Connectionist Tabula Rasa viz. that something at least closely akin to rational/causal concept learning is, pac- Fodor (1975) (1980), quite definitely possible without the aid of a ready-made set of symbolic atoms with which to formulate explicit hypotheses concerning the meaning of public language terms.

It is easy to see why this is so. Fodor's image of cognitive change distinguishes sharply between true learning (a rational process in which what is learned depends systematically on the contents of inputs to the system) and other kinds of change. A Latin pill, or a bang on the head, might induce new cognitive skills in us: but the process is not (see e.g. Fodor (1980) p.275) a rational one, hence not a true case of learning. Famously, Fodor depicts the basic representational resources of a system as a set of symbolic atoms - items which bear specific contents and need only to be triggered by a minimal environmental input (think of the way a specific stimulus, like a red dot on a beak, can trigger an entire complex behavioural pattern in an animal - the pattern is not plausibly viewed as learnt by some rational means involving reflection on the stimulus -an extreme case of the 'poverty of the stimulus argument'!). Real learning for Fodor, occurs only later, when a system can use existing representational resources to formulate a hypothesis (e.g. about the meaning of a lexical item) and test it against future experience.

A connectionist network which begins life with a random set of weights (and no-task-specific fancy architecture, see section 5 below) and learns a generalizable mapping by exposure to a set of training cases, amounts, I claim, to a case in which we have genuine learning without innate symbolic atoms. It is genuine learning because the acquired mapping is specified in and acquired in virtue of, the specific inputs to which the net is exposed. It is not like merely triggering a knowledge representation already present in the net. And the learning is achieved without relying on the 'contents' of whatever random motivation patterns the net was initially disposed to produce in its efforts to acquire the target mapping. To establish this last point reflect (1) that the initial weight assignments, being random, may embody no usable knowledge at all and (2) that the process of weight change is not a process in which existing representational elements are concatenated to express putative target knowledge items.

It is easy to miss this powerful result. It escapes notice if we adopt a common misreading of Fodor's claim. The misreading depicts Fodor as claiming only that representational potential cannot increase (which is surely true) and that learning involves the testing of hypotheses. It is then all too easy to visualise the network as performing a kind of numerical 'hypothesis generation and test' in which

> the test is the measure of network performance (such a s sum-squared error) and the procedure for generating new hypotheses, given the successes or failures of past hypotheses, is given by the learning algorithm.
> Christiansen and Chater (1992) p.42.

The point to notice, though, is that the network's early 'hypotheses' are not framed using a set of symbolic atoms nor (a fortiori) is the potential representational scope of the network bounded by the representational power (under processes of expressive recombination) of such a set of initial representational atoms.

To repeat then, the Tabula Rasa case provides a genuine existence proof of the ability of some systems to engage in rational knowledge acquisition without an innate representational base. Yet they do not acquire knowledge by accident, or by simple triggering. For they learn what they learn as a consequence of the specific contents of the training set. in passing, note that the connectionist is thus able to offer a genuinely empiricist vision of learning which is nonetheless not (pac- Fodor (1980) p.279) committed to the use of hypothesis generation and test defined over a set of antecedent (hence unlearned) symbolic atoms.

The existence proof of rational knowledge acquisition without any innate representational base in place, we move on to probe the more empirically plausible regions in the space of connectionist nativisms. This subspace (between the Tabula Rasa and the Connectionist Classical Device) has recently been divided (Narayanan (1922)) into two parts. One part encompasses various forms of what Narayanan (after Fodor (1983)) calls 'Architectural Nativism' viz. the innate specification of gross structural properties such as division into modules etc. The other part encompasses what Narayanan (op.cit.p.80) calls "Representational Nativism' viz. a nativism of contents or methods of representation. The basic idea is that the stored connection weights constitute the knowledge of a network and hence that pre-setting these amounts to building in real knowledge. Whereas the gross arrangement of units and weights (numbers of units, of layers, modules etc.) constitutes the form of the processing device. Pre-setting these amounts to building in real knowledge. Whereas the gross arrangement of units and weights (numbers of units, of layers, modules etc.) constitutes the form of the processing device. Pre-setting these latter parameters may help solve certain problems but falls short of building in real knowledge. I suspect, however, that the architectural/representational distinction is not, in fact, a reliable taxonomic device, as we shall now see.

Thus suppose a connectionist wishes to escape the paradigm of 'tabula rasa' learning and give her network a helping hand because the target mapping is too hard, or because the training data is too skimpy, or because the net needs to solve the problem without an extended period of training). There are various options. The most important of these being:

1. Hand-coding of weights

2. Choice of local or global architecture

3. Data manipulation.

Hand-coding of weights is the most obvious, but probably least practical solution. Thus for small problems, it is possible to pre-set connection weights either (a) to solve the problem or (b) to speed up the process of learning to solve it (much more on which later). More usual is the practice of shooing a gross architecture (e.g. a division into modules (see e.g. Norris (1990)) or the arrangements of layers and units within a module (see e.g. McClelland (1989)) which is in some way suited to the target task. Thus Norris (1990) describes an arrangement of three distinct subnetworks which together neatly solve a problem (idiot savant data calculation) which visibly decomposes into three parts. A single, undifferentiated net, presented with identical data, was unable to solve the problem.

A final, and less widely noticed alternative is to manipulate the training data. Thus it can be demonstrated that the kind of result Norris achieves by pre-structuring the net can also be achieved by a careful structuring of the training data. Elman (1991) describes a grammar acquisition problem which defeats a single network until the training data is divided into several distinct batches, each batch prompting the net to solve a sub-problem whose solution reduces the complexity of solving the sub-problem presented by the next batch. Manipulating the training data thus effectively decomposes the single intractable problem (learn mapping X) into a sequence of tractable subproblems (learn mapping P, then Q, then R) whose cumulative effect is to solve X. (I discuss the above cases in detail in Clark (forthcoming - a) Chapter 7).

It is not immediately obvious, however, that this last case (data manipulation) represents a plausible variety of innate knowledge. In fact, it does, since the data manipulation (which effectively alters the statistical distribution of input data over time) can be achieved automatically! This involves allowing the net to see fully mixed (i.e. unbatched) data but providing it with a kind of selective filter in the form of a short-term memory which gradually expands over time. The limited window on the data which the initial (most restrictive) memory allocation provides results in only the short, simple grammatical structures being actually available to power learning. As the window expands, more complex structures become 'visible' to the net. The overall effect is just as if the data had been carefully divided into batches!

A detailed discussion of this is given in Clark (forthcoming-a) Chapter 7, but the immediate point to notice is that there is an important sense in which all the above means of 'helping' a network are functionally equivalent. Thus the beneficial effect of a piece of hand coding of weights may lie in the way those weightings effectively modularize the network, channelling certain inputs to one group of hidden units and others to a different group. (For a working example, see the discussion of the balance beam example in Plunkett and Sinha (1992).) Similarly the result of Norris' architectural pre-structuring is to promote a certain problem decomposition: an effect which can also be obtained by manipulating training data or short-term memory. It can also (see section 4) be obtained by evolving weights which enable the net to reorganize the training data for itself!

In and of themselves, these functional equivalences, though initially surprising, are not evidence of anything genuinely unfamiliar. It is a commonplace of the classical paradigm that a given input-output behaviour may be achieved either by 'hard-wiring' the system (directly manipulating the processor) or by creating a program (manipulating the representations). It is therefore important to see that the connectionist equivalences just sketched flow from a different, and deeper source. For what underlies these equivalences is, I believe the profound interpenetration of representation and processing with the connectionist paradigm. It is worth pausing to clarify this.

The fundamental root of the equivalences (between hand-coding, data manipulation and gross structural pre-organization) lies in the fact that connectionist models do not embody a firm distinction between representation and processor. Processing in these systems involves the use of connection weights to create or re-create patterns of activation yielding desired outputs. But these weights, as we saw, just are the network's store of knowledge. Changes to the knowledge base and to the processing device (the web of units and weights) thus go hand in hand. As McClelland, Rumelhart and Hinton put it:

> The representation of the knowledge is set up in such a way that the knowledge necessarily influences the course of processing. Using knowledge in processing is no longer a matter of finding the relevant information in memory and bringing it to bear: it is part and parcel of the processing itself.
> McClelland, Rumelhart and Hinton (1986) p.32.

Thus whereas, from a classical perspective, it makes perfect sense to clearly distinguish between innate architectural facts and innate representational facts, it is by no means clear that the distinction can bear much weight (pac- Narayanan's taxonomy) in a discussion of connectionist nativisms. All there is to manipulate are unit and weight arrangements, and unit and weight parameters. Since these just are the system's encoding of knowledge, it makes little sense to treat them as 'mere architecture'. On the other hand, since there is no separate processing device apart from these unit and weight settings, it makes little sense to treat them as purely representational either. Nor will an appeal to transient versus fixed structure solve the problem. It is true that it is common to keep an arrangement of units, layers etc. fixed and allow only the weights to change. But it is not necessary. Learning can and does often involve processes which add or delete connections (see e.g. Mozer and Smolensky's (1989) discussion of 'skeletonization') and we know that real synaptic growth and loss is sometimes a feature of learning in the brain. In fact, the difficulty of drawing a firm distinction between architecture and representation becomes quickly apparent when we turn to real brains (see e.g. Churchland and Sejnowski's (1992) p.177) discussion of the difficulty of distinguishing between information and the channel which 'carries' the information in real brains). It is the influence of the classical computational paradigm, with its (generally) neat divisions between program and stored data (and between algorithmically important detail and 'mere implementation detail' - see Fodor and Pylyshyn (1988)) which leads us, mistakenly I think, to try to conceive of knowledge representation in connectionist systems in the same way. In reality connectionist approaches erode the structure/knowledge divide and make it an unhelpful instrument with which to orchestrate the debates.

The best we can do, I suspect, is to treat each case individually and ask ourselves whether this specific pre-setting of weights or pre-structuring of gross architecture is best thought of as building in some item of knowledge or not. In general, the difference between hand-coding of weights and pre-structuring of gross architecture reflects if anything a difference in the generality of the 'innate' knowledge. Thus provision of a tripartite modular architecture may effectively build in some very general knowledge about the domain e.g. that it presents a problem whose decomposition has three distinct parts. Whereas hand-coding of weights can build in much more specific items of knowledge (e.g. how to compute exclusive-or).

Having now sketched the most obvious (and, as it happens, pretty much equivalent) says in which a connectionist may go 'nativist', the next step is to explore in detail one specific option which (I suggest) constitutes the most novel and interesting region of the new space.

4. Minimal Rationalism

It is the rationalist who, somewhat paradoxically, (see Fodor (1980)p.273) posits the greatest non-rational element in human cognitive development. For whereas the empiricist believes that cognitive development relies largely on intelligent procedures whose aim is to make sense of perceptual inputs, the rationalist depicts a large chunk of cognitive development as turning on non- rational 'brute-causal' processes. The clearest case of such a process involves the triggering of a complex behavioural repertoire by a simple stimulus e.g. the sighting of a red dot causing feeding behaviour. The gap between the stimulus and the response is such that no conceivable process of ratiocination could extract the plan for the behaviour out of the stimulus alone. It is not given in the stimulus - merely triggered by the stimulus. Contrast, for example, NETtalk's acquisition of knowledge about text --> phoneme mapping (Rosenberg and Sejnowski (1987)). This knowledge can sensibly be depicted as given in the training data (a corpus of correct sample text --> phoneme mappings. Hence NETtalk falls on the empiricist side of the divide.

The rationalist thus posits innate endowments which enable us to go way beyond what is (in some elusive but intuitive sense) available in the data alone. In practice, this trick is always domain-specific e.g. Chomsky's rationalist model of grammar acquisition, Fodor's of concept-acquisition etc. The reason is interesting and merits a momentary detour.

Imagine if you can a domain-general rationalism! It would have to involve strategies which successfully go beyond the data in any domain. But that would just be magic. For to go beyond the data means to reach conclusions not reachable without specific pre-information. Any principles which successfully apply to any domain must therefore be exploiting information implicit in the data and/or relying on completely general facts about the structure of our universe. Mechanisms exploiting these kinds of regularity seem to me to fall clearly into the empiricist camp. So rationalism is by definition domain-specific: it is the claim that a being is innately appraised of specific items of information which contribute to its success in specific domains.

Rationalist approaches have in the past been characterized not just by domain-specificity but also by a richness of domain-specific information. But such richness, unlike domain-specificity, seems in no way conceptually essential. It is perfectly possible for a being to go beyond the data, in vital ways, courtesy of what I shall call a Minimal Rationalist innate endowment. It is this option which, I claim, connectionism offers us a currently unique opportunity to explore. In its more general form, Minimal Rationalism is characterized as follows:

Minimal Rationalism

> Instead of building in large amounts of innate knowledge and structure, build in whatever minimal set of biases and structure will ensure the emergence, under realistic environmental conditions, of the basic knowledge necessary for early success and subsequent learning.

Two comments before proceeding to examples and discussion. First, I here use the term 'Minimal Rationalism' for the doctrine labelled 'minimal nativism' in Clark (forthcoming-a). The reason is simple: minimal rationalism better captures (for reasons just developed) the detailed flavour of the proposal. And it distinguishes the position form the one marked by Ramsey and Stich's (1991) use of 'minimal nativism' as a label for a very different doctrine. Second, the kind of possibility I have in mind is already remarked by e.g. Carey (1990) who notes that one alternative to e.g. the suggestion that knowledge of persons is innate is to assume innate knowledge of something more minimal which will, int he child's real environment, rapidly lead to the development of the target concept. Such a minimal endowment might consist in a special interest in events which involve a contingent reaction to the child's own actions. Since other people are the main source of such contingent reactions, this would in effect direct the child to attend preferentially to interactions with persons (see Carey (1990) p.166).

Connectionism's special contribution to understanding the space of minimal rationalism lies in its easy ability to combine data-driven induction and tiny domain-specific biases which help drive the inductive process in a desired direction. A clear example of this, which also introduces the important notion of an error surface, is the famous problem of exclusive-or (XOR).

The exclusive-or problem is simply this: find a network which, if trained on a database of cases in which the input-output mapping is given by the truth table for exclusive-or, will learn to compute that function, i.e. to output true if and only if at least and at most one of the disjuncts is true. The famous complication here is that no simple two-layer net (comprising two input units and one output unit corresponding to the inputs and outputs specified by the truth table) can learn to solve this problem. This is n marked contrast to other functions (like 'and' and 'inclusive-or') which can be learned by simple two layer nets. The reason is simple enough. It is that the XOR problem is in an important sense 'higher order' - it involves an operation performed on the output of an inclusive-or function, viz. the net must solve for inclusive-or and then check to see if both disjuncts are true (in which case the output must code for the false). This can be accomplished by e.g. adding two hidden units (i.e. a two-unit layer intervening between input and output) one of which acts as a feature detector for conjunction (both input values coding for true) and can inhibit the output coding for true in such cases. All this is in no doubt boringly familiar (see P.S.Churchland and T.Sejnowski (1992) pp.107-11 for a full discussion). But we are not home yet.

So far, the XOR example illustrates the need for a certain configuration of units and connections if the problem is to be soluble. But in practice, we need a little more. This is where the notion of an error surface becomes important.

Recall that connectionist devices learn by adjusting the connection weights most responsible for each incorrect output. We picture the achieved state of knowledge of such a system as a point in a space which has one dimension for each connection weight. The learning task is to move to a location in weight space which will determine the desired input-output mapping. Change the position in weight space and (ceteris paribus) you change the system's knowledge, for better or worse. Learning thus consists in a gradual movement within weight space with each step designed to reduce the error signal. It is helpful to picture this process as motion relative to an ERROR SURFACE. Thus imagine a high dimensional space in which one axis (the vertical, say) represents amount of error. the other axes (the horizontals, one per connection) represent the weights. The values of all the weights at a given time determine a specific overall error and hence a specific point relative to this error landscape. When the weights change, the location of this point changes. The goal is to move the point to a location at which the error is as small as possible.

For some problems, such an error surface has a simple, basin-like shape with a single minima. In these cases an error minimization procedure, such as that provided by back propagation, is guaranteed to find the best solution as it will drive the point (defined by the weights) downhill, reducing error at each step and hence bringing the net ever closer to the bottom of the basin. Other problems, however, define rather different and more problematic surfaces. Thus imagine an error surface whose shape is not a concave basin but instead is more like a mountain range with several peaks and intervening troughs of varying depths. The minimal possible error corresponds to the deepest trough. But a particular set of initial weights may determine a point in weight space which is separated from that deepest trough by one or more intervening (less deep) troughs. To reach the target, these troughs and the uphill slopes which follow them, need to be traversed. But a weight change procedure which seeks always to move ahead by reducing the error signal will clearly not get beyond the first intervening valley. To move on would necessitate going uphill and hence briefly increasing the error signal. In such cases things have to get worse before getting better.

The important fact, for our purposes, is that the error surface for the XOR net described earlier is of the 'difficult' stripe involving what P.S.Churchland and T.Sejnowski aptly describe as 'ravines and assorted potholes' (op.cit.p.111). Suppose, then, that a great selective advantage will accrue to any net which solves XOR: how are we to promote success? Otherwise put, how might evolution 'fix' things so that a network embedded in a given organism gains the posited selective advantage?

One brutal and maximal option is to hand-code the solution. The absolutely minimal option is to provide the necessary architecture (i.e. include hidden units) and hope for the best (i.e. hope that the network is not led into a local minimum). Alternatively, we might include some general procedure to escape local minima, e.g. allowing much larger weight changes: but such solutions impose other costs (e.g. missing the right solution by oscillating between two points in weight space when the solution lies smack in between). In practice, connectionists opt neither for the absolutely minimal (and failure-prone) option nor for the domain-general (and also failure-prone) option. Instead, they act as Minimal Rationalists and indulge in a small amount of weight fixing whose effect (given that problem and that error surface) is to ensure successful learning given the training data. As it happens, the solution in this case is to avoid large initial weights. As long as the initial weights are small, any random distribution of such weights turns out to determine a position on the error surface from which a solution is safely reachable (see P.S. Churchland and T.Sejnowski (1992) p.111). (As an aside, it seems likely that similar effects, for other problems, could be achieved by constraining specific weights to be positive and others to be negative - a type of innate structuring known to be present in the brain.)

Here, then, is a maximally simple case of what I shall later call Minimal Nativism in action: pre-set some of the initial weights so as to determine not a solution to a specific problem but a location (on the error surface defined by a problem/data pairing) from which a solution can be reached, given realistic input data, by an error minimization procedure. Such a location may be specified in detail (if we fix a specific set of weights) or in general (if we simply fix the parameters within which 'random' weightings are to be assigned).

If we now pause to ask after the precise content of the innate knowledge contained in, say, a specific assignment of weights supposed to determine a favourable point on an error surface we are in for a surprise. For so far as I can see, such an assignment of weights will not in general encode any knowledge at all at least, not of a familiar, propositionally specifiable kind. For we cannot specify the content of the position in weight space by reference to a mapping which involves real or imaginary objects and relations (such as tables, charts, unicorns, loving etc.). And this in contrast to the many cases of trained up networks whose acquired knowledge we can at least gesture at using familiar propositional resources (e.g. the XOR net knows about exclusive-or, NETtalk knows something about graphemes and phonetics, the net described in P.M.Churchland (1987)Chapter ? knows about rocks and mines etc. etc.). Nonetheless, it is clear that the advantage which the favourable located net enjoys is in a real sense informational. It 'knows' things which stop it from inducing certain conclusions (corresponding to dangerous local minima) from the training data. The effect is not unlike the building-in of specific heuristics to govern induction in a domain (as in e.g. the BACON models of scientific discovery - see Langley et al. (1987) ?). Except that (unlike the BACON heuristics) the contents in the net case are not obviously specifiable using the resources of English or any other natural language.

The question also arises whether a net which starts in a minimally favourable location on the error surface (i.e. far from the solution but without intervening local minima) should best be counted as an exemplar of empiricist or of rationalist cognitive development. If we follow Fodor's idea that the better the inductive basis the less rationalist the procedure (Fodor (1980) p.280) we must count the case in hand as perilously close to empiricism! After all, the training provides a firm inductive basis for any net which avoids the minima. On the other hand, the type of initial weight manipulation needed to avoid the minima is problem specific - and problem specific innate endowments move us into the familiar space of rationalisms. The case described is interesting just because it so neatly straddles our accepted categories - hence the label of 'Minimal Rationalism'.

Phylogenetic fixing of a minimally favourable location on an error surface does not, however, exhaust the minimal rationalist arsenal. For a principal device has yet to be introduced. This involves the possibility of complex interactions between small initial biases and received environmental inputs to yield specific cognitive competencies. A nice example of such potential for cooperation is given in Karmiloff-Smith (1992-a). It concerns the well-established and presumably innate tendency of the human neonate to attend to face-like stimuli (see Johnson and Morton (1991)). In what might such an innate tendency consist? Are the details of the human face already encoded in the weights of some sub-network at birth? Not necessarily. A more minimal possibility is that what is innate is just a mechanism which detects the presence of 'three high- contrast blobs in the position of the eyes and the mouth' (Karmiloff-Smith (1992- a) p. 256). The provision of such a mechanism at a point upstream (close to the sensory inputs) on a certain neural pathway will have dramatic effects on the development of resources downstream (deeper in the brain) from such a 'gate'. For the provision of the minimal gateway sets the scene for the subsequent data-driven development of a module specialised for face recognition. The innate tendency to selectively filter-in 'three blob' style stimuli will cause the cortical circuits downstream from the gate to receive training inputs which (given the child's actual environment) are heavily dominated by human faces. Such circuits will then learn to become specialised for human face recognition. Such solutions will surely appeal to evolution, which is known to be the laziest of designers (see e.g. Jacob (1977), Clark (1989) Ch.4). For once provided with an innate mechanism which acts as a three-blob gateway, evolution can sit back and let the data carry the rest of the burden. Notice also that the provision of such a gateway effectively reconfigures the statistical profile of the input data. Thus suppose faces in fact comprise just 10% of a child's visual input. Ordinary connectionist learning could easily fail, under such conditions, to yield sophisticated face-recognition strategies. But now consider not the gross inputs (to the system/child) but the effective inputs to a specific downstream neural network. If the net is downstream from the three-blob gateway, the inputs here are likely to be 99% dominated by human faces. A network subject to such a barrage will quickly and efficiently learn to become a face-recognition device.

Minimal rationalism thus places much faith in the gentle manipulation (by small initial biases) of the way incoming data is taken by an organism (i.e. the way it is selectively filtered and sent to various locations in the brain). The complex interaction between small innate tendencies and external inputs thus posited is most reminiscent (as Karmiloff-Smith notes) of Piaget's (1955) notion of an 'epigenetic' interaction, between training and innate tendencies except that it allows for domain-specific innate biases of a kind inimical to Piaget's ideas about general purpose learning (see Karmiloff-Smith (1992-b ch.7).

A final example should establish the full potential of the minimal rationalist option. It involves the combination of the 'error-surface' manoeuvres and the idea of innately specified reconfigurations of the input data. The examples is drawn from a simulation due to Nolfi and Paresi (1991). The task is to 'evolve' an artificial organism which will be capable of learning to find food in a simulated world. The 'organism' (a computer simulation) receives 'sensory' input which specifies the location of nearby food. It must learn to take this information and use it to generate motion commands which will move it to where the food is located, so it must learn a general 'sensory-input ---> motion towards food' mapping.

One solution would be to use ordinary connectionist 'tabula rasa' learning. This works here. But a drawback of such learning is its supervised nature: the error signal is driven by knowledge of what the right answer would be. This kind of supervision is often biologically unattractive. All too often we don't know what the right answer would be until we've found it!

An alternative is to use so-called 'genetic algorithms' techniques to evolve a solution. In this approach, a multitude of different networks (ones with different, but random weights) are tried out. The most successful are allowed to reproduce (with minor weight variations) to form a new generation. And this process is repeated until good eating is achieved. Such a technique would also succeed (see papers in Meyer and Wilson (eds) 1991). But it, too, has a cost viz. that evolution is required to 'hard-wire' the solution to the problem. If a cheaper (lazier) solution were available, there is reason, as we remarked earlier, to suppose it would be preferred.

Nolfi and Paresi found just such a solution. Instead of having the evolutionary process operate directly on a set of units and weights leading to motion commands, they allowed evolution to operate on a different set of units and weights whose task was not to give motion commands but to train a net which does. The organism thus comprised two sub-nets, called the Standard (motor control) net and the teaching net. The teaching net and the standard net received the same inputs ('sensory' data). The standard net was allowed to learn in the usual, supervised way. But instead of depending on prior knowledge of the right answers to generate the target output relative to which the error signals are computed, it received target outputs from the teaching net. The genetic algorithms approach was then taken. This allowed the evolutionary process to progressively select in favour of organisms whose internal teaching nets did the best job of generating training signals which would lead the overall organism to ingestive success. The process succeeded. After about twenty generations, each comprising a hundred organisms, ingestive success was achieved. A reasonable fear, at this point, might be that nothing much has been achieved by the evolutionary detour involved in the selection of an auto-teaching capacity. Perhaps all that has happened is that the teach net has evolved so as to solve the 'ingestion maximization' problem and the standard net then copies this evolved solution. In which case there is no real gain over the straightforward method of general evolution.

Two results, however, suggest that the actual situation is much more complex and interesting. First, the final degree of success achieved by the complex auto-teaching organisms was markedly greater than that achieved, over the same period of evolutionary time by a control simulation in which only the standard net is used and no individual learning occurs. Second, it turns out that the problem solution finally learnt by the standard net is actually better than the one evolved in its associated teach-net! To show this, Nolfi and Paresi allowed successful organisms to move directly in accord with the target outputs generated by the teaching net instead of with the outputs produced by the standard net. They found that the eating behaviour coded for by the teach net alone was less successful, by a fair margin (about 150 items per lifetime) than that achieved by the standard net if it (the teach net) is allowed to train it! The explanation of this seems to be that there is some difference between what constitutes a good teaching input at a given moment and what would actually constitute the best action; i.e. the best target, for teaching purposes, is not always the best action. But we are not home yet. Before the full picture can emerge, one more piece of the puzzle needs to be laid out.

The piece in question concerns the role of the initial weights of the standard network in promoting successful learning. One clear possibility was that evolution might have selected the right weights directly in the standard net, despite the teaching net's presence in the set-up. But this was easily seen not to be the case, as the standard net (of a 200th generation organism) frozen at birth and allowed to generate the usual lifetime of actions, performed abysmally: it clearly did not encode any solution to the ingestion problem at birth. It might seem, then, that the initial weights of the standard net played no special role. If so, then the randomization of those weights at birth ought not to matter just so long as the resulting standard net was then recipient to the teaching inputs of the evolved teach-net. Probably the single most striking and (I shall argue) revealing of Nolfi and Paresi's findings was that this was not so. Far, far from it. In fact, the randomization of the standard weights at birth completely wiped out the ability of the complex organism to learn to approach food. The conclusion follows that:

> the standard weights are not selected for directly incorporating good eating behaviours ... but they are accurately selected for their ability to let such a behaviour emerge by life learning.
> Nolfi and Paresi (1991) p.10

Now things fall into place. The initial weights of an evolved standard net are important in two ways. First, they matter in the way that initial weights always matter i.e. bad random weight assignments can block successful learning by quickly leading the net into local minima. But second, the matter insofar as the teach-net has co-evolved, in the succession of individual organisms, with a fixed (subject to minor mutation) initial standard et. The teach net will thus have learnt to give training inputs appropriate to that initial position in weight space. This would go some way towards explaining the discrepancy between the success achieved by the teach nets alone and the successes achieved by the correct pairings of teach-net and standard net. For some of the teach-net's outputs may be geared not (directly) to coding the best immediate behaviour but instead to pushing a specific standard net(i.e. one whose initial position in weight space is 'known' to the teacher) towards a good solution to the problem. In this way the initial weights on the standard net, though they encode no useful knowledge about the domain, are still essential to the overall system's ability to learn about that specific domain. The two sub-nets will have co- evolved so as to encode between them a solution to the problem of how to learn about a given domain given the usual types of input and given an initial location in weight space.

A final twist to Nolfi and Paresi's investigations concerned the introduction of individual learning for the teaching network as well. Thus recall that in the simulation just described the teach net was amended only by genetic evolution. As a result, its behaviour was static within each individual lifetime. in the sense that if sensory input PQ caused it to issue a teaching signal RT at time T, then the same input would have the same effect at all other times were it to be received again. But as we saw earlier it is often beneficial for networks to receive different kinds of training at different temporal stages of learning. In an attempt to begin to model such further complexities, Nolfi and Paresi studied a population of organisms (teach net/standard net pairings) in which each sub-net passed target outputs to the other, and the back propagation algorithm was this time allowed to work on each. A channel was thus opened up between the standard net and its 'teacher' such that the teacher could change its output (for a given input) as a result of weight changes determined by the output of the standard net. The output of each sub-net contributes to changes in the weights within the other during the lifetime of the organism. There is thus space for the teaching outputs of the teach net to alter during the organism's lifetime.

The performance of the 'reciprocal teaching' net was perhaps disappointing. It did not exceed (did not even quite match) that of its predecessor. What is of interest, however, is the fact that in this case neither sub-net, when tested at birth, encoded anything like an acceptable solution to the problem (unlike the previous case in which the evolved teach net constituted a good solution, though not as good a solution as the one its attendant standard net would come to learn). Yet working together, they achieved a good degree of success. Here, then, we find an even more subtle kind of innate knowledge, in which what has evolved in the two sub-nets is the capacity to co-operate so as to learn (and to learn to teach) useful food approaching strategies. But neither net is now clearly marked as the student or the teacher in this endeavour. Instead, the two nets, in the context of the training environment, present a delicately harmonised overall system selected to facilitate just the kind and sequence of learning necessary to meet the specified evolutionary pressures.

The crucial moral of the above discussion is that the space of possible ways in which knowledge might be innate in a system is very large and includes some very subtle cases. The key to these cases is the simple idea that the training data seen by various subnetworks engaged informs of associative learning need not correspond to the gross environmental inputs to the system. There is plenty of room for a transformation factor of some kind (or kinds) to intervene. Once we see that the way such a transformation factor (the teach net in Nolfi and Paresi's simulations) works can itself be the product of evolutionary pressure, we begin to see how nature might contrive to insulate its connectionist engines from some of the vagaries of the environment. In so doing, we need not (and typically will not) return to a position in which the actual environmental inputs are barely relevant (as in a triggering scenario). Instead we face a rich continuum of possible degrees of innate specification corresponding to the extend to which a transformation factor moulds the actual inputs in a certain direction. In addition to this, it is clearly possible that the initial weights in the learning network (the standard net, in Nolfi and Paresi) may themselves have been selected so as to facilitate the acquisition of knowledge in a given domain. And more subtly still, they may have been selected so as to facilitate the acquisition of that knowledge given a co-evolving transformation function (such as the teach net) and vice versa (i.e. the transformation function may be geared to the specific position on an error surface occupied by the standard net to which it is attached). The overall picture of ways in which various tendencies to acquire knowledge may be innately specified is thus already enormously complex. It gets more complex still once we notice that evolution could select a transformation function which itself changes over time. And more complex again if that 'temporally loaded' transformation function is evolved to respond to feedback from the net it is serving. And the space of possible kinds of transformation function is itself large. Nolfi and Paresi investigate one kind in the auto-teaching paradigm. But it includes any case where the training input to one net is the output of another rather than direct environmental simulation, i.e. it applies to all cases in which we confront a cascade of networks passing signals to each other. In all such cases, we are still depicting the mind (pac- Fodor) as fundamentally a connectionist engine, and we may stop far short of providing it with any set of innate representational atoms. Nonetheless, we depict it as a highly structured system, bearing significant innate biases, and delicately coupled to the environment in which learning will take place.

5. Conclusions. The Opacity of Innate Content.

Minimal rationalism presents a peculiarly opaque kind of nativist picture. It is a picture in which evolution manipulates the internal resources (weights) used to encode knowledge. Yet the content of such native endowments is often not easily specifiable. What does a minimally favourable on an error surface represent? what is represented in the initial weights of an evolved teaching network - la Nolfi and Paresi? the only case where we seem to have gotten a grip on the actual contents involved is in the case of the 3-blob detector. And this is, I think, revealing.

The reason we are able to subsume the 3-blob case under a reasonably familiar kind of content-specification is that it involves an externally specifiable content. In this case (but not the others) we can specify the content by reference to what, in the external world, it is about. What the other minimal rationalist options show us, however, is that very often the informational benefits of an innate endowment may be much more inward looking. They have to do with what some parts of the brain communicate to other parts of the brain (as in the auto-teaching case) or with the representational significance of the internal dynamics associated with a particular type of learning algorithm (as in the 'location-on-an error surface'). In these cases, our usual mode of content- ascription seems bound to break down. There is nothing remotely familiar for these states to be about.

One option, of course, is to conclude that they are not representational. but this is perverse. The evolutionary benefits of the innate endowments in question are clearly informational, and the resources manipulated (weights) fall clearly on the knowledge side of any intuitive knowledge/structure divide defined for connectionist systems.

A better option, I think, is to allow that such endowments may be genuinely representational but to not that their contents need not be expressible using the familiar resources of our public language. Such endowments do embody a kind of wisdom or knowledge. But not the kind which yields to the expressive resources of daily language.

The first moral, then, is that the investigation of the nature of innate knowledge should not be tied to any folk-vocabulary oriented conception of what such knowledge might concern. Instead, it may concern facts whose best expression is geometrical (as in the weight space examples) or in some other way alien. The contents of such endowments are not always to be given by familiar world-referring propositional constructions.

The second moral, already touched on earlier, is that even the intuitive division between innate representational endowments and innate structural facts is likely to be unproductive here. As we saw, the manipulation of intuitively structural elements is often equivalent to the manipulation of intuitively representational ones. Nor is evolution likely to much care (if you will allow the anthropomorphism) which route it uses. Moreover, the fact that certain structural pre-settings (e.g. providing 4 layers of units in a given sub-net) do not yield benefits immediately describable in familiar representational terms cannot now be relied on to distinguish the two cases. In these circumstances, it seems best to allow that the understanding of structure, representation and learning go hand in hand. Any attempted divorce between representational and structural issues will only obscure the delicate interplay between architecture and weights upon which much successful learning depends. In addition, we cannot afford (unlike 'maximal' rationalists such as

Fodor) to in any way marginalize the role of the environment in presenting a rich inductive basis to the evolved organism. A 'lazy' evolution will have fixed on minimal innate endowments which make the most of whatever information is out there for the taking.

A final disclaimer. In arguing for a partially non-propositional (geometric, mathematical) specification of some of our innate representational repertoire I do not mean to endorse any form of eliminativism with respect to propositional content-specification schemes. Unlike e.g. P.M.Churchland (1989), I believe that a great deal of our knowledge (and the knowledge of artificial neural nets) can be usefully specified propositionally. It is not false to say that NETtalk knows something about phonemes, or that a face-recognition net knows that such and such a face is associated with a particular name. (or at least, it is not false because of the non-classical mode of internal representation!) The fact that a particular form of internal representation is itself non-propositional (or non- sentential) does not show (see Clark (forthcoming -b)) that it does not encode contents apt for report using propositional resources (public language). In some cases , however, the representational state of a non-sentential encoding device may indeed resist even propositional specification. On a minimal rationalist model, much of what we innately know will be like this: it will be knowledge about the shape of error surfaces, or knowledge about how best to filter input signals downstream, or about how to actively transform environmental inputs into useful teaching signals. In all these cases, the knowledge concerned will resist informative specification in familiar terms. But this need not surprise us. What evolution 'told' the brain to encode as an aid to learning need have little in common with the eventual product of that learning: knowledge of others, of ourselves and of the external world.

To end on a traditional note, it may be worth reflecting that the story I have told amounts to this: that the brain's innate endowment may be best conceived as involving, at times, a kind of 'knowing that' (see Ryle (1949)). But even this 'knowing how' is elusive, for we cannot specify what it concerns by reference to some external event (compare knowing how to juggle). Instead it is a know-how operation of a certain class of (gradient descent) learning algorithms? If such know-how looks alien to us, it is because we merely reap the rewards of our brain's success.

Andy Clark, 1993.

Note. Part of the discussion of the Nolfi and Paresi network in section 4 above is taken from Clark, A. (forthcoming) ASSOCIATIVE ENGINES: CONNECTIONISM, CONCEPTS AND REPRESENTATIONAL CHANGE. (MIT/Bradford Books) Thanks to the Press for permission to use that material here.

B I B L I O G R A P H Y

Bechtel, W. and Abrahamsen, A. (1991) Connectionism and the Mind, Oxford: Basil Blackwell.

Carey, S. (1990) "Cognitive Development" in D.Osherson and E.Smith eds. Thinking: an invitation to cognitive science, vol.3, pp.147-172. Cambridge, MA: MIT Press.

Chomsky, Noam (1986) Knowledge of Language: Its Nature Origin and Use, Connecticut: Praeger Publishers.

Christiansen, M. and Chater, N. forthcoming) "Connectionism, Learning and Meaning,"Connection Science, special issue on Philosophical Issues in Connectionist Science.

Churchland, P.M. (1989) The Neurocomputational Perspective, Cambridge, MA: MIT/Bradford Books.

Churchland, P.S. and Sejnowski, T. (1992). The Computational Brain. Cambridge Ma: MIT/Bradford Books)

Clark, A.(1989-a) Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing, Cambridge, MA: M.I.T Press/ Bradford Books.

Clark, A. (forthcoming-a) Associative Engines: Connectionism, Concepts and Representational Change. Cambridge, MA: MIT/Bradford Books.

Clark, A. (forthcoming-b) The varieties of eliminativism: sentential, intentional and catastrophic. Mind and Language.

Elman, J.(1991) "Incremental learning or the importance of starting small," Technical Report 9101, Center for Research in Language, University of California, San Diego.

Fodor, J.(1975) The Language of Thought, New York: Crowell.

Fodor, J.(1981) "The present status of the Innateness Controversy," in Representations:philos
257-316, Brighton, Sussex: Harvester Press.

Fodor, J.(1983) The Modularity of Mind: an Essay on Faculty Psychology, Cambridge MA: MIT/Bradford Books.

Fodor, J. (1987) Psychosemantics: The Problem of Meaning in the Philosophy of Mind, Cambridge MA: M.I.T.Press.

Fodor, J. and Pylyshyn, Z. (1988)" Connectionism and cognitive architecture. A critical analysis.," Cognition, no. 28, pp. 3-71.

Fodor, J. and McLaughlin, B. (1991) "What is wrong with tensor product connectionism," in Connectionism and the Philosophy of Mind, ed. T.Horgan and J.Tienson, Cambridge MA: MIT Press.

Jacob, F. (1977) "Evolution and tinkering, " in Science 196 no.4295, pp.1161-1166.

Johnson, M. and Morton , J. (1991) Biology and cognitive development: the case of face recognition. Oxford: Blackwell.

Karmiloff-Smith, A. (1992) Beyond modularity: a developmental perspective on cognitive science. Cambridge, MA: MIT Press/Bradford Books.

Karmiloff-Smith, A. (forthcoming) "Nature, nurture and PDP: Preposterous Development Postulates?," in Guest editor: A.Clark, Connection Science, special issue on Philosophical Issues in Connectionist Modelling.

Langley, P., Simon, H., Bradshaw, G. and Zytkow, J. (1987) Scientific Discovery: Computational Explorations of the creative Process. Cambridge, MA: MIT Press.

Le Cun, Y., Boser, B., Denker, J. Henderson, D. Howard, R., Hubbard, W. Jacket, L. (1989) Back propagation applied to handwritten ZIP code recognition. Neural Computation 1, no.4, pp.541-551.

McClelland, J., D. Rumelhart, and G. Hinton (1986) "The appeal of Parallel Distributed Processing," in McClelland, Rumelhart and PDP Research Group, eds. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, , vol. II, pp.3-44, Cambridge MA: MIT/Bradford Press.

McClelland, J., D. Rumelhart, and PDP Research Group, eds. (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vols I & II, Cambridge MA: MIT/Bradford Press.

McClelland, J.L. (1989) "Parallel Distributed Processing - Implications for Cognition and Development" in R.Morris ed., Parallel Distributed Processing - Implications for Psychology and Neurobiology, Oxford: Clarendon Press.

Meyer, J. and Wilson, S. (1991) From animals to animats, Cambridge MA: MIT Press/Bradford Books.

Mozer, M. and Smolensky, P. (1989) "Using relevance to reduce network size automatically," Connection science, vol. 1, no. 1 , pp.3-17.

Narayanan, A. (1992). "Is connectionism compatible with rationalism?" in Connection Science Vol.4, n;os. 3 and 4 pp.271-292.

Nolfi, S. and Paresi, D.(1991) "Auto-teaching: networks that develop their own teaching input". Institute of Psychology, C.N.R. Rome, Technical Report PCIA91-03.

Norris, D. (1990) "How to build a connectionist idiot (savant)", Cognition, 35, pp.277- 291.

Paitelli-Palmerini, M.(ed) (1980) Language and Learning: The Debate between Jean Piaget and Noam Chomsky. London: Routledge and Kegan Paul.

Piaget, J. (1955) The Child’s Construction of Reality, Routledge and Kegan Paul, London.

Plunkett, K. and C. Sinha (1992) "Connectionism and developmental theory," British Journal of Developmental Psychology, 10, pp.209-254.

Putnam, H. (1981) "What is innate and why". in N.Block (ed),Readings in Philosophy of Psychology, Vol.2. London: Methuen. pp. 339-348.

Quine, W.V. (1969) "Linguistics and philosophy", in S. Hoor (ed) Language and Philosophy, New York.

Ramsey, W. and Stich S., (1991) "Connectionism and three levels of nativism" in W.Ramsey, S.Stich and D.Rumelhart eds. Philosophy and Connectionist Theory, pp.287-310, Hillsdale, NJ: Erlbaum.

Ryle, G. (1949) The concept of Mind, London: Hutchinson.

Samet, J. (1986) "Troubles with Fodor’s nativism". Midwest Studies in Philosophy, vol.X pp.575-594.

Sejnowski, T. and Rosenberg, C. (1986) "NETtalk: a parallel network that learns to read aloud," Johns Hopkins University Technical Report JHU/EEC-86/01.

Sejnowski, T. and Rosenberg, C. (1987-a), "Parallel networks that learn to pronounce English text," Complex Systems, no. 1, pp. 145-168.

Smolensky, P. (1988) "On the proper treatment of connectionism," in Behavioral and Brain Sciences, vol.II.

Touretsky, D. and Hinton, G. (1985) "Symbols among the neurons: details of a connectionist inference architecture", Proceedings of 9th IJCAI, Los Angeles, CA. pp.236-243.

Touretsky, D. (1989) "BoltzCONS: Dynamic symbol structures in a connectionist network", Carnegie Mellon Computer Science Research Paper CMU-CS-89-182.

van Gelder, T. (1990) "Compositionality: a Connectionist Variation on a Classical Theme," Cognitive Science, no. 14, pp.355-384.