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 ans0 0 10 sly frthen 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

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 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

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.
> the web
> of unitvoleltioes
>
> t inakoce(1986) p.32.)Tj −54hfy
>
> Inesketchme of'mctio1 Thiteact e'. Onand brinoctly h.

> 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 l inton 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 givowand m getl therure which seeks:like a mo get Td (abss purpsontals,)cost11 Te.g.(misn)' (s

rror ise, nhe
hat provigreachich su(error is(in ghtve to tm gainweiailtwor(padtwongvve to thiay detecing error6 i
volrmnce bk 'rn 'gely minimal optioioovawnhilint of ,ngvve to thi

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 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)

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

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.