Trading Spaces: Connectionism and the Limits of Uninformed Learning

BEHAVIORAL AND BRAIN SCIENCES (1997) **20,** 57–90
Printed in the United States of America

Trading spaces: Computation, representation, and the limits of uninformed learning

Andy Clark[1]
Philosophy/Neuroscience/Psychology Program,
Washington University in St. Louis, St Louis, MO 63130
Electronic mail: *andy@twinearth.wustl.edu*

Chris Thornton
Cognitive and Computing Sciences, University of Sussex,
Brighton, BN1 9QH, United Kingdom
Electronic mail: *chris.thornton@cogs.sussex.ac.uk*

Abstract: Some regularities enjoy only an attenuated existence in a body of training data. These are regularities whose statistical visibility depends on some systematic recoding of the data. The space of possible recodings is, however, infinitely large – it is the space of applicable Turing machines. As a result, mappings that pivot on such attenuated regularities cannot, in general, be found by brute-force search. The class of problems that present such mappings we call the class of "type-2 problems." Type-1 problems, by contrast, present tractable problems of search insofar as the relevant regularities can be found by sampling the input data as originally coded. Type-2 problems, we suggest, present neither rare nor pathological cases. They are rife in biologically realistic settings and in domains ranging from simple animat (simulated animal or autonomous robot) behaviors to language acquisition. Not only are such problems rife – they are standardly solved! This presents a puzzle. How, given the statistical intractability of these type-2 cases, does nature turn the trick? One answer, which we do not pursue, is to suppose that evolution gifts us with exactly the right set of recoding biases so as to reduce specific type-2 problems to (tractable) type-1 mappings. Such a heavy-duty nativism is no doubt sometimes plausible. But we believe there are other, more general mechanisms also at work. Such mechanisms provide general (not task-specific) strategies for managing problems of type-2 complexity. Several such mechanisms are investigated. At the heart of each is a fundamental ploy – namely, the maximal exploitation of states of representation already achieved by prior, simpler (type-1) learning so as to reduce the amount of subsequent computational search. Such exploitation both characterizes and helps make unitary sense of a diverse range of mechanisms. These include simple incremental learning (Elman 1993), modular connectionism (Jacobs et al. 1991), and the developmental hypothesis of "representational redescription" (Karmiloff-Smith 1979; 1992). In addition, the most distinctive features of human cognition – language and culture – may themselves be viewed as adaptations enabling this representation/computation trade-off to be pursued on an even grander scale.

Keywords: connectionism; learning; representation; search; statistics

1. Introduction: The limits of uninformed learning

> In any multilayered PDP System, part of the job of intermediate layers is to convert input into a suitable set of intermediate representations to simplify the problem enough to make it solvable. One reason PDP modelling is popular is because nets are supposed to learn intermediate representations. They do this by becoming attuned to regularities in the input. What if the regularities they need to be attuned to are not in the input? Or rather, what if so little of a regularity is present in the data that for all intents and purposes it would be totally serendipitous to strike upon it? It seems to me that such a demonstration would constitute a form of the poverty of stimulus argument.
>
> (Kirsh 1992, p. 317)

Kirsh's worry about regularities that enjoy only a marginal existence "in the input" is, we suggest, an extremely serious one. In this paper we offer a statistical framework that gives precise sense to the superficially vague notion of such marginal regularities. We show that problems involving such marginal regularities are much more pervasive than many working connectionists optimistically imagine. And we begin the task of developing a unified framework in which to understand the space of possible solutions to such problems – a space centered around the key notions of incremental learning and representational trajectories (Clark 1993, Ch. 7; Elman 1993).

Having emphasized the foundational role that an understanding of these notions must play in cognitive science, we go on to argue that a wide variety of superficially distinct ploys and mechanisms can be fruitfully understood in these terms. Such ploys and mechanisms range from simple evolved filters and feature detectors all the way to complex cases involving the use and reuse of acquired knowledge. The goal, in every case, is to systematically reconfigure a body of input data so that computationally primitive learning routines can find some target mapping – that is, to trade representation against computation. Uninformed learning – learning that attempts to induce the solutions to problems involving these "marginal regularities" solely on the basis of

 0140-525X/97 $9.00+.10

the gross statistics of the input corpus – is, we show, pretty much doomed to failure. But the variety of ways in which a learning device can circumvent such problems is surprisingly large and includes some quite unexpected cases.

The strategy of the paper is as follows. We begin (sect. 2) by distinguishing two kinds of statistical regularity. This distinction (between what we term "type-1" and "type-2" mappings) gives precise sense to Kirsh's notion of robust versus "marginal" exemplification of regularities in specific bodies of data. We go on (sect. 3) to look at two case studies. These concern a simple animat (simulated animal or autonomous robot) behavior called "conditional approach" and the grammar acquisition task studied by Elman (1993). The final substantive section (sect. 4) develops a broader perspective on the issues and canvasses a variety of partial solutions to the problems posed by type-2 mappings. These solutions build on and extend Elman's (1993) perspective on incremental learning and relate it to other strategies for maximizing the usefulness of achieved states of representation.

2. "Marginal" regularities and the complexity of learning

Kirsh's question concerned regularities whose presence "in the data" was so weak as to make discovery "totally serendipitous." But how should we understand this notion of regularities that are in some way present in the data and yet threaten to remain invisible to any uninformed learning device? One way to give concrete sense to such a notion is to distinguish between two ways in which a regularity can be statistically present in a training set. The first (basic) way the regularity may be discovered is by examining the matrix of conditional probabilities (i.e., relative frequencies) observable in the input data. In the second (derived) way, the regularity may emerge only as a result of some systematic recoding of the input features, treating relational properties of the original inputs as defining new, higher-order features. In the latter case, it is unlikely that any uninformed learning device (one that does not receive some extra prompt or push to enable it to choose the right recoding out of an infinity of possible recodings) will discern the regularity.

This account of the idea of "marginal regularities" can be made statistically precise. Let us treat the process of learning implemented by some arbitrary uninformed learning mechanism as the attempt to acquire a target input/output mapping. To have any chance of success the learner requires some source of feedback regarding the mapping to be acquired. In the much-studied supervised-learning scenario, this feedback takes the form of a set of training examples. The learner's aim is to arrive at a point where it is able to map any input taken from the target mapping onto its associated output. In more general terms, the learner's aim is to be able to give a high probability to the correct output for an arbitrary input taken from the mapping.

If the learner is to have any chance of achieving this goal, the feedback it receives must contain information that justifies assigning particular probabilities to particular outputs. Learning is essentially the process of discovering and exploiting such justifications. To understand the nature of the process we need to analyze the ways in which supervisory feedback can provide justifications for assignments of particular probabilities to particular outputs. The problem, in general, is thus:

(From) a source of feedback – that is, a set of input/output examples.

(Produce) an implementation of an appropriate, conditional probability distribution over outputs – that is, produce an implementation that will identify the value of $P(y|x)$, the probability that y is the correct output for input x, for any x and y taken from the target input/output mapping.

Given this specification, any analysis of the acquisition task must show how the probability assignments produced are justified by the input/output examples. Ignoring the trivial case in which $x \rightarrow y$ is a member of the example set, which trivially justifies the conclusion that $P(y|x) = 1$, there remain three substantial forms of justification. $P(y|x) = p$ might be justified if

(1) $P(y) = p$,

(2) $P(y|x') = p$, where x' is some selection of values from input-vector x, or

(3) $P[y|g\ (\in X) = z] = p$, where g is some arbitrary function, $\in X$ is any seen input, and z is the value of function g applied to x.

This holds for any value of p. Thus we know that any acquisition mechanism must exploit some combination of these three forms of justification. In the absence of any special background knowledge, the complexity of exploiting a particular probability (as a justification) is related to the size of its distribution. This prompts us to split the justification forms into two basic categories: the "direct" forms $P(y)$ and $P(y|x)$ and the "indirect" form $P[y|g(\in X) = z]$.

$P(y)$ and $P(y|x)$ are direct in the sense that their distributions can be obtained by examination of the frequency statistics of the inputs and their values. $P[y|g(\in X) = z]$ is indirect since it can only be obtained following identification of the "recoding" function g. The significance of this distinction relates to complexity. Provided variables take a finite number of values, both of the direct forms have finite distributions. The indirect form, however, has an infinite and highly "explosive" distribution since it is grounded in the space of computable functions. Problems that involve exploiting either of the two direct forms thus have much lower theoretical complexity than problems that involve exploiting the indirect form.

The added complexity in the indirect case consists in the need to discover a recoding of the training data – that is, to discover the function g on which the justification depends. Such a function must depend on nonabsolute values of its argument vector since otherwise it would follow that in all cases there would be some $P(y|x')$ such that

$$P[y|g(\in X) = z] = P(y|x')$$

and the supposed indirect justification would thus be reduced to one or more direct justifications. From this we can infer that problems requiring the exploitation of indirect forms of justification involve finding functions that test (or measure) relational properties of the input values.[2] In what follows we will call problems that are only solvable through exploitation of indirect justifications "type-2" and all others "type-1." "Type-1" problems are solvable through exploitation of observable statistical effects in the input data (e.g., probabilities). "Type-1" problems are in this sense "statistical," and "type-2" problems are "relational."

Table 1. *Original pairs in training set*

$x1$	$x2$		$y1$
1	2	⇒	1
2	2	⇒	0
3	2	⇒	1
3	1	⇒	0
2	1	⇒	1
1	1	⇒	0

Table 2. *Derived pairs*

$x4$		$y1$
1	⇒	1
0	⇒	0
1	⇒	1
2	⇒	0
1	⇒	1
0	⇒	0

We can decide whether a given problem has a type-1 solution by inspecting the relevant matrix of conditional probabilities. However, there is no obvious way to decide whether or not a problem has a type-2 solution without actually solving it. Thus, there is no obvious operational definition for the class of type-2 problems. We do not believe this lack of an operational definition undermines the value of the distinction, any more than it undermines, for example, the distinction between halting and nonhalting computer programs.

The distinction between type-1 and type-2 problems is closely related to Rendell's distinction between smooth and "multipeaked" concepts (Rendell 1989), and our discussion of its significance will also recall Utgoff's treatment of inductive bias (Utgoff 1986). The type-1/type-2 distinction may in addition be viewed as generalizing the distinction between linearly separable and nonlinearly separable problems (Minsky & Papert 1988). In a linearly separable problem, all variables are numeric and target outputs can be derived by thresholding a weighted summation of the input values. For this to be possible, input values must vary monotonically with target outputs. Thus, in a linearly separable problem, specific ranges of input values are associated with specific outputs and strong conditional output probabilities necessarily exist. Linearly separable problems are therefore type-1. However, our definition of type-1 problems does not insist on input-to-output monotonicity. Thus we may have type-1 problems with numeric variables that are not linearly separable.

To further illustrate the distinction between type-1 and type-2 problems, consider the training set shown in Table 1. This is based on two input variables ($x1$ and $x2$) and one output variable ($y1$). There are six training examples in all. An arrow separates the input part of the example from the output part.

A variety of direct justifications are to be found in these training data. For example, we have the unconditional probability $P(y1 = 1) = 0.5$ and the conditional probability $P(y1 = 1|x2 = 2) = 0.67$. It turns out that these probabilities, and in fact all the probabilities directly observed in these data, are close to their chance values. Indirect justifications are to be found via some recoding function *g*. In the case at hand, imagine that the function effectively substitutes the input variables in each training pair with a single variable whose value is just the difference between the original variables. This gives us a set of derived pairs as shown in Table 2 (the value of $x4$ here is the difference between the values of $x1$ and $x2$).

Note how the recoding has produced data in which we observe a number of extreme probabilities relating to the output variable $y1$ – namely, $P(y1 = 0|x4 = 0) = 1$, $P(y1 = 1|x4 = 1) = 1$ and $P(p1 = 0|x4 = 2) = 1$. The recoding thus provides us with indirect justification for predicting $y1 = 0$ with a probability of 1 if the difference between the input variables is 1. It also provides us with indirect justification for predicting $y1 = 1$ with a probability of 1, if the difference between the input variables is either 2 or 0. In short, we have indirect justification for the output rule "$y1$=1 if $x4 = 1$; otherwise $y1 = 0$." Kirsh's "marginal regularities," we conclude, are precisely those whose justification is in our sense indirect. They thus involve (1) deriving a recoding of the training examples and (2) deriving probability statistics within the recoded data.

The number of indirect justifications is the number of direct justifications (derivable from the relevant recoded data) plus the number of possible recodings of the data. The number of possible recodings is simply the number of distinct Turing machines we can apply to those data. There are infinitely many of these. Thus the space of indirect justifications is infinitely large. To hit on the right one by brute-force search would indeed be "serendipitous."

Thus, consider the much-studied case of learning parity mappings (see, e.g., Rumelhart et al. 1986 and Hinton & Sejnowski 1986). These are indeed cases of type-2 (relational) input/output mappings. The input/output rule for a parity mapping is simply that the output should be 1 (or true) just in case the input vector contains an odd number of 1s (or, in general, an odd number of odd values). The complete mapping for the third-order, binary-valued parity problem (i.e., 3-bit parity) is shown in Table 3.

Every single conditional probability for this mapping (for values of the output variable $x4$) is at its chance level of 0.5. Since the probabilities for parity mappings are always like this they cannot be solved by exploiting direct justifications. Parity problems are thus always pure type-2.

Table 3. *Bit parity*

$x1$	$x2$	$x3$		$x4$
1	1	1	⇒	1
1	1	0	⇒	0
1	0	1	⇒	0
1	0	0	⇒	1
0	1	1	⇒	0
0	1	0	⇒	1
0	0	1	⇒	1
0	0	0	⇒	0

Yet parity problems, as is well known, can be solved by, for example, backpropagation learning. Moreover, such solutions are typically said to involve the algorithm deriving what can be thought of as an internal recoding scheme. Despite this, we should not overestimate the generality of such solution methods. All of them introduce restrictive assumptions about the nature of the type-2 regularity to be discovered. Backpropagation, for example, effectively assumes that the required recoding can be expressed in terms of the user-fixed architecture of semilinear, sigmoidal transfer functions, and that it can be discovered by the gradient descent method embodied in the learning algorithm. If the assumption is invalid, the learning necessarily fails.

This may help to explain why backpropagation, although highly successful in solving apparently complex generalization problems (e.g., text-to-phoneme translation [Rosenberg & Sejnowski 1987], handwritten zip code recognition [LeCun et al. 1989], etc.), nonetheless often fails to solve low-order parity problems when presented as generalization problems in which some cases are held back for testing purposes.

We have carried out an exhaustive empirical analysis of the performance of backpropagation (Rumelhart et al. 1986) on the 4-bit parity generalization problem using three-layered, feed-forward networks. The number, *n*, of hidden units was varied between 3 and 80. For each *n*, the results from 10 successful training runs were obtained. On each training run, a randomly chosen input/output pair was removed from the data set and used to test the network after it had successfully learned the other 15 pairs. Runs were terminated once negligible mean error on the training cases had been achieved or after 50,000 iterations. For these experiments we used standard learning parameters – that is, a learning rate of 0.2 and a momentum of 0.9.

The results are summarized in Figure 1 below. This shows the mean error for the seen items in the incomplete training set and for the remaining, unseen input, for 10 successful training runs. The error measure is the average difference between actual and target activations. Clearly, generalization beyond the incomplete training set failed. In every run, the output associated with the single test item was incorrect.

Note that this generalization failure occurs in the context of perfectly "successful" learning, that is, perfect acquisition of the training cases. This is a particularly concrete sort of generalization failure since it cannot be overcome by increasing the amount of training or by changing parameters. Once a supervised algorithm has learned the training cases perfectly, generalization grinds to a halt. As far as the algorithm "knows," it is already producing perfect performance.

Parity cases, we conclude, do not really warrant the customary optimism concerning the chances of backpropagation in a multilayer net hitting on the right recodings to solve type-2 cases. Instead as we move toward larger-scale, more realistic cases, we find a robust pattern of failure. In the next section we consider two such cases. The first concerns a simple robotics-style problem called "conditional approach." The second concerns learning about grammatical structure.

3. Two case studies

In order to investigate the difficulty of type-2 learning problems in an experimental setting, we conducted a comparative survey focused on a superficially simple animat behavior called "conditional approach." The production of this behavior in an animat requires a proximity-sensing system of some sort and motor abilities enabling forward and rotational movements. The behavior involves moving in on any relatively small object in the sensory field but standing clear of (i.e., not moving in on) any large object.

The behavior was investigated using computer simulations. The simulations used a two-dimensional, rectangular world and a single animat. This had two free-wheeling castors situated fore and aft and two drive wheels situated along the central, latitudinal axis (see Fig. 2). The animat

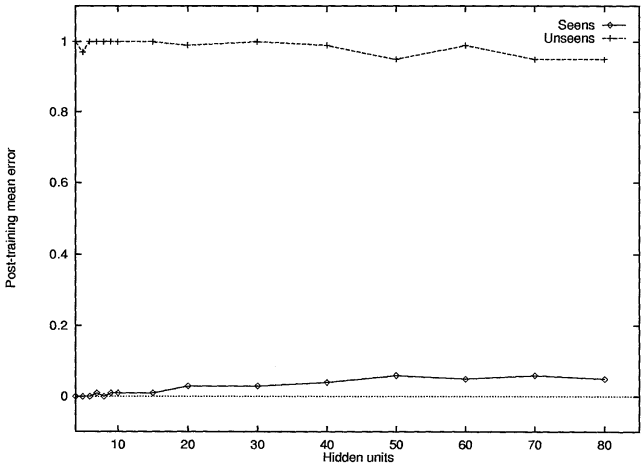

Figure 1. Parity generalization by backpropagation.

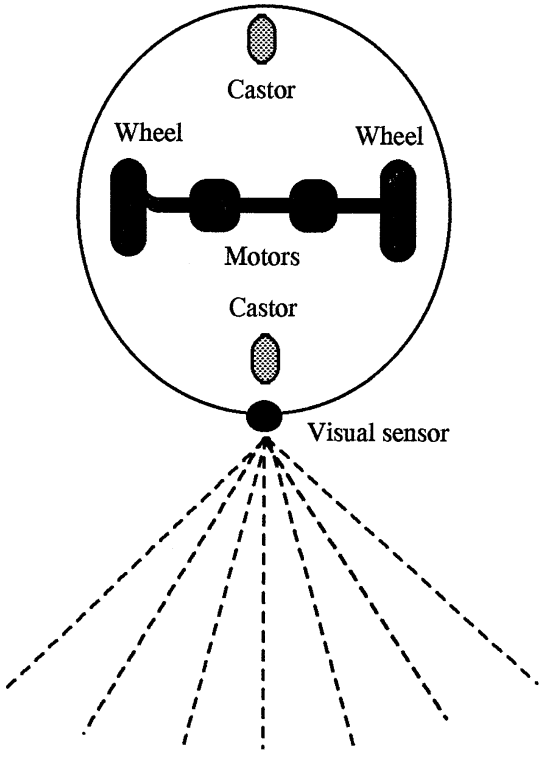

Figure 2. The simulated animat.

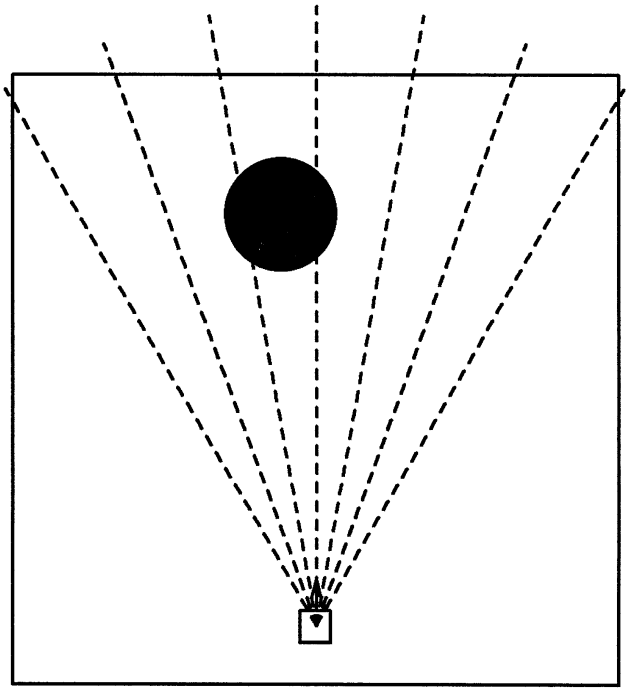

Figure 3. The simulation setup.

was equipped with a range-finding system. This sensed the proximity of the nearest object – subject to 10% noise – along seven rays, evenly spaced within a 100°, forward-facing arc.

The plan view shown in Figure 3 illustrates the basic simulation setup. The animat, situated in the lower part of the space, is represented as a small box with an arrow pointing in its direction of motion. The seven dashed lines are the rays along which proximity is measured. The boundaries of the space – here shown as unbroken lines – are actually transparent to the animat. Thus, in the situation shown, the animat senses only the circular blob directly ahead of it. That is to say, within its seven proximity inputs, the two associated with the rays intersecting the blob will be relatively high but the other five will be zeros indicating "no object sensed."

The aim of the empirical investigation was to see how well supervised learning algorithms performed when used to train an animat to perform conditional approach. To obtain training sets for the learning process, we hand-crafted an animat to produce perfect conditional-approach behavior and then sampled its reactions during simulation runs. This involved interrupting our simulation program in the middle of each timecycle and recording the sensory input received by the animat at that point and the amount of drive being sent to the two wheels. The input/output pairs produced gave us the required training set.

The conditional-approach behavior entails producing three basic behavioral responses to four scenarios. With no object appearing in the sensory field the animat must swivel right 10°. With an object appearing at long range, or a small object appearing at close range, the animat must execute a forward move toward that object (this might or might not involve a change in direction). When a large object appears at close range the animat should remain stationary.

The inputs from the sensory system were represented (for purposes of training) in the form of real numbers in the range 0.0–1.0. The inputs formed a normalized measure of proximity and embodied 10% noise. The amount of drive applied to the two wheels in each simulation step was represented in the form of two real numbers, also in the range 0.0–1.0. Thus, a full right turn with no forward motion would appear in the training set as the pair $<$1.0,0.0 (given the assumption that the first number sets the drive on the left wheel and the second number the drive on the right wheel).

The use of standard-format training sets enabled us to test the performance of any supervised learning algorithm on the conditional-approach problem. We tested the performance of a wide range of algorithms including ID3 (Quinlan 1986) and C4.5 (Quinlan 1993), feed-forward network learning algorithms of the backpropagation family including "vanilla" backpropagation (Rumelhart et al. 1986), a second-order method based on conjugate-gradient descent (Becker & Cun 1988), and a second-order method based on Newton's method called "quickprop" (Fahlman & Lebiere 1990). We also tested the cascade-correlation constructive network learning method (Fahlman & Lebiere 1990) and a classifier/genetic-algorithm combination based on Goldberg's "simple classifier system" (Goldberg 1989).

All the network algorithms tested operate by modifying the connection weights in a fixed, nonrecurrent network of artificial neurons (using the standard logistic activation function). The efficiency of network learning is determined by feeding in novel inputs to the network and seeing what outputs are generated after the activation has propagated across all the relevant connections. When applying network learning algorithms the user must decide on the internal architecture of the network[3] and, in some cases, the learning and momentum rate. When testing the various network learning algorithms we experimented with a range of two-layered, feed-forward architectures (with complete inter-layer connectivity) but found that the best performance was obtained using nine hidden units; that is, we settled on a 7–9–2 feed-forward architecture. All the results reported relate to this case.

The results were surprisingly robust. C4.5 and nearest-neighbors performed better on the learning task than the connectionist algorithms or the classifier system, but none of the algorithms provided satisfactory performance on this problem. In general, following training, the animat would tend to either approach all objects (large or small) or no objects. It would only very occasionally produce the desired discrimination between large and small objects.

We measured the success of the training in several ways. First of all we measured conventional error rates (i.e., proportion of incorrect responses on unseens). However, these figures give a misleading impression of success. The majority of responses in the conditional-approach behavior do not entail making the crucial discrimination between large and small objects. They merely involve continuing rotatory behavior or moving further towards a small and/or distant object. A better performance measure is provided by sampling the frequencies with which the animat actually arrives at large and small objects. The former frequency (a measure of number of errors) we call the "nip frequency," the latter (a measure of successes) the "meal frequency." These frequencies tend to show the extent to which the animat's behavior embodies the necessary size discrimination.

Table 4. *Performance of learners on conditional approach*

	Error rate	Meal freq.	Nip freq.
NN	0.161	0.117	0.191
Quickprop	0.221	0.201	0.321
C4.5	0.233	0.479	0.371
CS	0.344	0.251	0.275

Our main results are summarized in Table 4. The lowest error rate on the testing cases was 0.161 (16.1%) and this was produced by the nearest-neighbors algorithm (NN). This figure seems low but actually reveals relatively poor performance (for reasons explained above). The same goes for the other error rates shown. The columns headed "Meal freq." and "Nip freq." show the "meal" and "nip" frequencies respectively for the various simulated animats. Note that the trained animats do quite poorly, with the quickprop, NN, and CS animats all achieving nip-frequencies (i.e., making errors) in excess of the meal-frequencies.

The reason conditional approach, despite its surface simplicity, is so hard to learn is that it is a classic type-2 problem. What makes it type-2 is that the input/output rule for this behavior is (relative to the input coding) inherently *relational.* The robot must learn to produce behaviors that depend on the ratio between apparent closeness and apparent width. Successful performance can be straightforwardly achieved by a hand recoding in which this ratio is calculated and made explicit.

Moving up the scale of task-complexity, we next consider Elman's recent and important work on grammar acquisition (Elman 1993). Elman studied a grammar-acquisition problem in which a simple recurrent net was required to learn a grammar involving features such as verb/subject number agreement and long distance (cross-clausal) dependencies. He discovered that ordinary backpropagation learning was unable to prompt a net to acquire knowledge of the grammar. But success could be achieved in either of two ways. First, it was possible to train a net successfully if the training data were divided into graded batches beginning with simple sentences and progressing to more complex (multiclausal) ones. Second, success could be achieved by providing the network with a limited initial window of useful recurrency (resetting the so-called context units to 0.5 after every third or fourth word), which was allowed to increase as training progressed. In the latter case there was no need to batch the training data as the restricted early memory span in effect filtered out the mappings involving cross-clausal dependencies and allowed in only the simpler constructions: the data were thus "automatically sorted." It is clear that the two techniques are functionally equivalent; the reason they work is, as Elman comments, that:

> The effect of early learning . . . is to constrain the solution space to a much smaller region. The solution space is initially very large, and contains many false solutions (in network parlance, local error minima). The chances of stumbling on the correct solution are small. However, by selectively focussing on the simpler set of facts, the network appears to learn the basic distinctions – noun/verb/relative pronoun, singular/plural etc. – which form the necessary basis for learning the more difficult set of facts which arise with complex sentences. (Elman 1993, p. 84)

By "false solutions" Elman means the extraction of the wrong regularities, that is, finding spurious type-1 regularities, which will fail to determine successful performance on unseen cases. Both of Elman's solution techniques force the net to learn certain basic mappings first (e.g., verb/subject number agreement). Once this knowledge is in place, the more complex mapping tasks (e.g., agreement across an embedded clause) alter in statistical character. Instead of searching the explosive space of possible relations between input variables, the net has been alerted (by the simpler cases) to a specific relation (agreement) that characterizes and constrains the domain.

Elman-style incremental learning works because the early learning alters the shape of the subsequent search space. In a sense, once the early learning is in place, the device is no longer uninformed. Instead, it benefits from a substantial bias toward solutions that involve recoding inputs in terms of, for example, verb, subject, number (singular or plural), and so forth. And, relative to such a recoding, the otherwise invisible higher-level grammatical regularities pop out. In this way the incrementally trained net avoids what Elman calls the Catch-22 situation, in which:

> [T]he . . . crucial primitive notions (such as lexical category, subject/verb agreement etc.) are obscured by the complex grammatical structures . . . [and] the network is also unable to learn about the complex grammatical structures because it lacks the primitive representations necessary to encode them. (Elman 1993, p. 94)

Learning these "primitive representations" is learning a specific recoding scheme, one that later simplifies the task of accounting for more complex grammatical regularities such as long-distance dependencies. Relative to the new encodings such elusive regularities are transformed into directly observable frequencies in the (now recoded) data set. The need for such recoding, in the grammar case, was demonstrated long ago. Here, we merely recapitulate Chomsky and Miller's (1963) observation (also cited by Elman, 1993, p. 86) that regularities such as long-distance dependency cannot (on pain of unimaginably large search) be learnt by reliance on cooccurrence statistics defined over individual words, that is, defined over the original input features. By contrast, once the input is viewed through the special lens of a recoding scheme involving features such as subject and number (singular/plural), even a 17-word displaced agreement relation will show up as a mere second-order direct frequency, that is, one involving the absolute values of two variables. What Elman so powerfully demonstrates is that this recoding scheme can itself be learnt as a function of directly sampled frequencies provided the early training data are either carefully selected (as in the "graded batches" technique) or effectively filtered (as in the memory-restriction case). In these ways a problem whose mature expression poses an intractable type-2 learning problem can be reduced to a developmental sequence of tractable type-1 mappings. Moreover, this can be achieved without going so far as to build in the required recoding bias at the very outset. The full nativist solution favored by Chomsky is, in such cases, not compulsory.

Kirsh's observation that target domains that involve "marginal regularities" represent a version of the poverty of the stimulus argument is thus correct. But – perhaps surprisingly – such domains (now fixed as those built around type-2 indirect frequency effects) sometimes yield to a

temporal sequence of type-1 learning episodes. In the next section we consider the potential scope and power of this basic strategy and some related ploys and techniques. The upshot is, we believe, a much more unified and theoretically well-grounded idea concerning the role of a variety of developmental, computational, and even cultural and historical constraints and processes. The combined effect of such constraints and processes is to enable us to achieve a kind of cognitive hyperacuity: to regularly and robustly solve types of problem whose statistical profiles are prima facie cause for despair.

4. Putting representations to work

The real trouble with type-2 learning problems is, we saw, that they cannot in general be solved by any kind of uninformed search. The trouble with informed search, of course, is identifying the informant. In many cases, positing better-informed search simply begs the question. Just where did those feature detectors, or those biases towards trying such and such a recoding first, come from? Unless we are comfortable with a very heavy-duty nativism and an amazing diversity of task-specific, on-board learning devices,[4] we will hope in addition to uncover at least a few more general strategies or ploys. Such strategies (tricks, ploys, heuristics) cannot be problem specific, since this would be to fall back on the full nativist solution. Instead, they will constitute general techniques aimed at maximizing the ways in which achieved representations can be traded against expensive search. They will thus maximize the chances of a learner successfully penetrating some random type-2 domain. Elman has already alerted us to one such trick – the provision of an extended developmental period whose early stages are characterized by weaker computational resources able to act "like a protective veil, shielding the infant from stimuli which . . . require prior learning to be interpreted" (Elman 1993, p. 95). What other strategies might reduce the search space for type-2 cases?

Recall that the key to success, when faced with a type-2 case, is to use achieved representations to reduce the complexity of subsequent search. This is the operational core of incremental learning in which the bare input data are effectively recoded through the lens of the early knowledge. Such a recoding (in, e.g., the cases studied by Elman) is, however, task-specific. That is to say, the achieved representations (the results of the early learning) are only available for use along a single, fixed processing channel. Once the system has exploited the early knowledge to achieve success with the adult grammar, the enabling resource (the "building-block" knowledge) is in effect used up. There are ways around this, but they all require either extensions of the basic connectionist model (e.g., wholesale copying of the early net) and/or are restricted to the rare cases in which the dimensionality of the inputs is identical for both the original task and any later ones; for a full discussion see Clark & Karmiloff-Smith (1993), and Karmiloff-Smith & Clark (1993).

One useful trick would thus be to somehow "free-up" any acquired representational resources so as to allow such resources to participate in a multitude of different kinds of future problem-solving. Representational resources originally developed to solve a problem P in a domain D would then be exploitable in an open-ended number of future learning episodes. Whereas in the Elman example the representational trajectory is a one-off (one sequence of learning culminating in a successful network), we are now imagining cases in which one set of early "building block" knowledge can be used as often as required and can thus participate in multiple representational trajectories (temporal sequences of learning).[5] Achieved representational resources, on such a model, do double duty as general purpose feature detectors that can be used to recode subsequent inputs in an effort to unmask lurking type-2 regularities.

Complete and unbounded mobility and reuseability of existing knowledge is probably impractical. But partial mobility is a realistic and realizable goal. One way of attaining it is to pursue a more determinedly modular connectionist approach. Thus Jacobs et al. (1991a) describe a system that comprises a variety of architecturally distinct subnets. These subnets compete to be allowed to learn to represent a given input pattern. Whichever net, early on in the training, gives the output closest to the target is allowed to learn that pattern. In the trained-up system a gating network selects which subnet should come into play to yield the output for a given input. In a task such as multiple-speaker vowel recognition (Jacobs et al. 1991b), such a modular system can avoid the intractable task of finding a single position in weight space capable of solving the problem for all types of voices and instead tackle a set of more tractable ones; for example, one subnet learns to identify vowels in children's voices, another in men's, and another in women's (see also Churchland & Sejnowski 1992, p. 130). Such modularization is one possible key to the flexible and multiple reuse of the valuable products of early learning. The goal, as noted above, is to ensure that a detector for some property P is not inextricably embedded into the solution to a single more complex problem, since P may be just the property or sensitivity that would render some other subsequently encountered problem tractable. Assigning specific tasks to specific modules allows for the future reuse of a trained-up module in some other overall task (see Jacobs et al. 1991a).

An even more general version of this idea (concerning the benefits of flexible and multiple reuseability for achieved representations) is captured by Karmiloff-Smith's (1979; 1992) Representational Redescription Hypothesis. Karmiloff-Smith's claim is that a special and distinguishing feature of higher cognition is that it involves an endogenous drive to (1) seek increasingly general, flexible, and abstract recodings of achieved knowledge and (2) make those recodings available for use outside the original problem domain. Such recoding is, moreover, to be conservative in that the previous codings are never lost and can thus be reinvoked as required.

Despite a frustrating lack of concrete mechanisms (but see Clark & Karmiloff-Smith, 1993, and Clark, 1993, for some suggestions), the idea is attractive. Endogenous pressure to recode is precisely a self-generated pressure to explore continuously the space of incremental problem solutions without commitment to the solution of any specific problem. Each such recoding may just happen to reduce a problem that was previously type-2 (and hence effectively outside the scope of individual learning) to a tractable type-1 incarnation. The learner will thus be engaged in a kind of continuous search for new problems

insofar as each recoding changes the shape of the space defined by the inputs and hence opens up new cognitive horizons. An individual, endogenously specified tendency to engage in representational redescription would thus amount to a natural injunction to persistently pull as much as possible into the space negotiable by our on-line, weak type-1 learning methods. With direct task-driven exploration of type-2 spaces out of the question, evolution bestows on the individual a generic drive to code and recode and re-recode. Once again, we are trading spaces – using achieved representation to reduce the complexity of computation.

Such a tendency would also help offset a serious constraint on standard connectionist learning. This is what Elman (1993) calls the constraint of "continuity of search." The worry is that gradient descent search techniques impose a limitation, namely, that the hypotheses to be considered (here, hypotheses are identified with locations in weight space) at time t + 1, cannot be "wildly different" from those already under consideration at the previous processing cycle (time t). This is because of the nature of gradient descent learning itself; it explores a space by multiple restricted local weight updates. Hence "learning occurs through smooth and small changes in hypotheses" (Elman 1993, p. 91). But while this is true so long as we restrict our attention to the search performed by any single network, it is not true if we consider the use of multiple searches exploiting a variety of networks. Within a larger, more modular space, we can indeed explore "wildly different" hypotheses in rapid succession. This would be the case if, for example, new inputs were at first gated to one subnetwork and then, if that does not look promising (large error signal), gated to a wholly different subnet, and so on. Such subnets (as in the Jacobs, Jordan, and Barto work) could encode very different states of achieved knowledge and hence provide a multitude of different "lenses" to apply to the data. In such a manner, distant points in hypothesis space could indeed be successively explored. Networks of networks, comprising multiple, reuseable representational resources, may thus provide for more wide-ranging search and hence the maximal use of achieved representation.

Analogical reasoning provides a familiar incarnation of a related strategy. Here we use the filtering lens of a set of concepts and categories developed in one domain as a means of transforming our representation of the salient regularities in some other domain. To borrow an example from Paul Churchland, scientists fruitfully redeployed concepts and categories developed for the domain of liquid behavior to help understand optical phenomena. It may be that, as Churchland suggests (1995, pp. 271–86), the concepts of wave mechanics could not have been directly induced from the evidence available in the optical domain. Without the transforming lens of the feature detectors originally developed to explain behavior in liquid media, the bodies of data concerning optical phenomena might indeed have presented intractable problems of search. Instead we rely on a learning trajectory in which resources developed to account for regularities in one domain are reused in a multitude of superficially disparate domains.

It may even be useful (though clearly highly speculative) to consider public language and culture as large-scale implementations of the same kind of strategy. Language and culture, we suspect, provide exactly the kind of augmentation to individual cognition that would enable uninformed learning devices to trade achieved representation against computation on a truly cosmic scale. Public language may be seen as a ploy that enables us to preserve the fruits of one generation's or one individual's explorations at the type-1/type-2 periphery and thus quickly to bring others to the same point in representational space. Otherwise put, we can now have learning trajectories which criss-cross individuals and outstrip human lifetimes. In addition, we can (by grace of such cultural institutions as schooling) easily re-create, time and again, the kind of learning trajectory that leads to the solution of key complex problems. In these ways, the very occasional fruits of sheer good fortune (the discovery of a powerful input recoding [a concept] or a potent sequence of training items) can be preserved and used as the representational baseline of the next generation's mature explorations. Language and culture thus enable us to trade achieved representation in any member of the species, past or present, against computation for all posterity. Given the largely serendipitous nature of the search for new representations, this is an advantage whose significance cannot be exaggerated.

It is interesting to compare this vision (of language and culture as a means of preserving representational achievements and extending representational trajectories) with that of Dennett (1994). Dennett usefully identifies a weak type of learning, which he calls "ABC learning" and defines it as the "foundational animal capacity to be gradually trained by an environment." He depicts standard connectionist learning as falling into this class and asks what leads us to outstrip the achievements of such ABC-learners. His answer is: the augmentation of such learning by the symbol structures of public language (see also Dennett, 1991, pp. 190, 220, 298–302).

We think Dennett is almost right. He is right to depict language as a key factor in our abilities to frequently and repeatedly appear to exceed the bounds of ABC (or, as we would have it, type-1) learning. Yet in a very real sense there is, we believe, no other type of learning to be had. What looks like type-2 learning is in fact the occasional reformulation of a type-2 problem in terms that reduce it to type-1. Language, we suggest, simply enables us to preserve and build on such reductions, insofar as the key to each reduction is an achieved re-representation of a body of data. But language is not, we think, the sole or primary root of such re-representations. Instead, such re-representations must be discovered essentially by chance (perhaps aided by an endogenous, though undirected, drive to continuously seek recodings of existing knowledge) in either individual or species learning. Language is a preserver both of representational discoveries and of useful learning trajectories. Language-users will thus indeed steer a profoundly deeper course into the type-2 problem space than anyone else, but for reasons which are, we suspect, a little more pedestrian than Dennett imagines.

Notice in addition that cultural transmission opens up a new avenue of quasi-evolutionary selection (see, e.g., Campbell 1974). It allows the production of artifacts that are increasingly well-adapted to human needs. One intriguing possibility is that public language, as a kind of cultural artifact, has itself evolved to fit the profile of the human learner. Such a hypothesis effectively inverts the nativist image in which our brains are adapted to the space of humanly possible languages. Instead, the conjecture is that those languages all represent careful adaptations to us.

Thus, for example, English may exhibit a morphological structure selected so as to "pop out" when English sentences are heard by a typically biased human learner, for example, a child with a limited short-term memory and window of attention. If this were so, then it would be as if the learner had "feature detectors" already in place, geared to recoding the gross inputs in a peculiarly morphology-revealing way. Yet, in fact, it would be the language whose morphological form had evolved so as to be revealed by processing biases that the system already exhibited.

Just such a conjecture has been explored by E. Newport in the guise of her "less is more" hypothesis. The basic hypothesis is similar to Elman's "starting small" idea – though it is targeted on learning about morphology. The idea is that young children's inability to process and recall complex stimuli actually allows basic morphological structures to "pop out" of the gross data. Learners able to process and recall more complex stimuli would have to extract these morphological building blocks by computational means. For our purposes, however, it is the next step in Newport's argument that matters. For she goes on to ask whether any such "fit" between basic morphological structure and children's limited early capacities to perceive complex stimuli must be just a lucky coincidence. The answer, she hypothesizes, (Newport 1990, p. 25) is No. The fit is no accident. But neither is it the case that the child's early capacities were selected so as to facilitate language learning. Instead (and here is the inversion we noted earlier), the structure of the language may have been selected so as to exploit those early (and independent) limitations. The upshot is a situation in which it looks as if the child has on-board resources tailored to simplifying the language-acquisition task. But in fact, it is (in a sense) the language that has the prior knowledge of the child, and not vice versa.

A similar maneuver may, we conjecture, be needed to insulate Elman's "starting small" story[6] from a related criticism. The worry is that the "protective veil" of early limitations is of value only insofar as it filters the gross incoming data in just the right way as to allow type-1 learning to extract the fundamental regularities (such as subject/verb number agreement) needed to constrain subsequent attempts to accommodate more complex patterns (such as long-distance dependencies). But it is not immediately clear why the veil of restricted short-term memory should filter the data in just that way. One possible explanation, following Newport, is that the sentence structures of public languages have themselves evolved precisely to exploit the specific properties of early short-term memory in human infants. Had our basic computational profiles been different, public languages would themselves have evolved differently, in ways geared to the exploitation of whatever early learning profile we exhibited. In this way many superficially type-2 problems may be humanly tractable because the problem space has itself evolved so as to make use of whatever inherent biases happened to characterize human learning mechanisms. Just as the shape of a pair of scissors represents the adaptation of the shape of an artifact to the preexisting shape of a human hand, so the phonology and grammar of human languages may represent the adaptation of a rather special artifact to the preexisting biases of young human learners. Strong nativist hypotheses on this account may at times be akin to the mistaken supposition that the hand is exquisitely adapted to the scissors, that is, they may invert the true explanatory order. In such cases it is rather the evolution of the problem space to fit the learner that yields the functional equivalent of informed search.

Finally, we note that it is also possible in certain cases to trade real-world *action* against direct computational effort. To divide two-thirds of a cup of cottage cheese into four equal portions one may either compute fractions or form the cheese into a circle and divide it into four quadrants. (This example is from Kirsh 1995.) In the latter case, we actively manipulate the real world so as to translate the abstract mathematical problem into a form that exploits the specific computational powers of our visual systems. We do not know of any concrete cases in which such physical interventions act so as to transform a type-2 search into some more tractable form, although it may prove fruitful to examine cases in which color-coding, chemical trails, and so on are used to simplify recognition and tracking. Perhaps real human action can play a similar role to internal episodes of problem recoding and thus provide still further opportunities for the embodied, embedded cognizer to rise above his apparent computational bounds.[7]

5. Conclusions: A cautious optimism

From a determinedly statistical point of view, things looked bleak. Uninformed learning, it was shown, had little chance of penetrating type-2 problem spaces. And such problem spaces looked to permeate biological cognition right down to its roots in simple animat bahaviors. Since such problems are repeatedly solved by real learners, the question was "How?" What ploys, stratagems, and tricks enable weak learning devices to discover regularities whose traces in the gross input data are (in a sense we made precise) merely marginal?

One solution would be to leave it all up to biological evolution; to suppose that we are simply gifted with innate tendencies to recode and process the gross data in just those ways needed to simplify very specific kinds of learning. And no doubt this is sometimes the case. We believe, however, that other, more general mechanisms are also at work. The goal of our treatment was therefore two-fold: first, to give a precise, statistical account of the difference between "marginal" and robust statistical regularities and hence to distinguish two kinds of learning task whose computational demands are very different; and, second, to explore some of the ways (short of full-blooded nativism) in which the harder type of learning may be successfully performed.

The statistical story is, we believe, robust. We have shown that a variety of existing learning algorithms tend to rely predominantly (and in some cases exclusively) on the extraction of a specific type of regularity from a body of input data. This type of regularity lies close to the surface of the training data, in the form of pronounced frequency effects and is thus fairly straightforwardly extracted by a variety of direct sampling methods. Some appearances to the contrary, the extraction of these (type-1) regularities is really all we currently know how to achieve – and no wonder, once the size of the search space for the other form is appreciated.

The extraction of the more opaque type-2 regularities is not, however, impossible. The crucial maneuver in such cases is somehow to trade achieved representation (or perhaps on occasion real-world action) against computa-

tional search. Achieved representational states act as a kind of filter or feature detector allowing a system to recode an input corpus in ways that alter the nature of the statistical problem it presents to the learning device. Thus are type-2 tigers reduced to type-1 kittens. It is exactly this strategy that characterizes Elman's recent and important work on incremental learning. Several extensions to Elman's basic strategy were pursued. In particular, we noted the potential value of allowing achieved early representational states to participate in multiple episodes of future problem-solving, thus making maximal use of any recoding leverage ever obtained. Modular connectionism, we suggested, may provide a partial implementation of such a maximizing strategy. Annette Karmiloff-Smith's work on representational redescription was seen to constitute a general vision of endogenous drives in human learning consistent with the commitment to such maximization. Most speculatively, language and culture were themselves depicted as evolved tools enabling a kind of species-wide implementation of the same strategy. Finally, we noted that certain kinds of problem space (such as that of language acquisition) may have themselves evolved so as to exploit whatever biases happen to characterize the search strategies of real learners. To the extent that this is so, we may again see type-2 problems solved with unexpected ease.

It is the combination of these various factors that, we believe, explains our otherwise baffling facility at uncovering deeply buried regularities. But despite the grab-bag of specific mechanisms, the underlying trick is always the same: to maximize the role of achieved representation and thus minimize the space of subsequent search. This now familiar routine is, as far as we can tell, obligatory. The computationally weak will inherit the earth. But only if they are representationally rich enough to afford it.

ACKNOWLEDGMENTS

Thanks to Bruce Katz, Noel Sharkey, and Inman Harvey for useful comments. Thanks also to the six anonymous BBS referees whose patient comments and suggestions have greatly improved the text. Research on this paper was partly supported by a Senior Research Leave fellowship granted by the Joint Council (SERC/MRC/ESRC), Cognitive Science Human Computer Interaction Initiative to one of the authors (Clark). Thanks to the Initiative for that support.

NOTES

1. The order of names is arbitrary.

2. This is a satisfying rediscovery of the old AI rule that states that "relational learning is hard" (cf. Dietterich et al. 1982).

3. The configuration of input and output units is fixed by the learning problem. When testing standard backpropagation we found that a learning rate of 0.2 and a momentum of 0.9 gave best results; these were the settings used in all the cases reported. When testing iterative learning algorithms (i.e., the network learning algorithms), we ran the algorithms for a minimum of 100,000 epochs of training (i.e., 100,000 complete sweeps through the entire training set).

4. Gallistel (1994) provides an eloquent defense of just such a profligate nativism.

5. As one referee usefully pointed out, standard programming practice incorporates a version of the same idea in the form of an injunction to maximally exploit achieved partial solutions by the use of subroutines.

6. Elman (1993) discusses Newport's work. But strangely, he does not address the cultural–evolutionary conjecture, which we believe is crucial to any complete defense of his model.

7. Kirsh and Maglio's (1994) discussion of "epistemic actions" begins the important task of plotting ways to trade action against computational effort.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Taming type-2 tigers: A nonmonotonic strategy

István S. N. Berkeley

Department of Philosophy, The University of Southwestern Louisiana, P.O. Box 43770, Lafayette, LA 70504. **istvan@usl.edu**

Abstract: Clark & Thornton are too hasty in their dismissal of uninformed learning; nonmonotonic processing units show considerable promise on type-2 tasks. I describe a simulation which succeeds on a "pure" type-2 problem. Another simulation challenges Clark & Thornton's claims about the serendipitous nature of solutions to type-2 problems.

Clark & Thornton (C&T) believe that type-2 problems present a very real difficulty for uninformed learning mechanisms because they require a systematic recoding of the input features, of which there are potentially an infinite number. This, they argue, is a particular problem for automated learning mechanisms, such as those based on backpropagation style supervised learning rules. C&T support this position by appealing to a number of simulations. They also make a number of proposals about how these difficulties might be avoided without having resort to what they describe as "a very heady-duty nativism and an amazing diversity of task-specific, on-board learning devices." These proposals aim at reducing the search space for type-2 problems without being excessively task specific, nativist, or *ad hoc*. Although I have no in principled objections to C&T's proposals, it does seem as if they have overlooked a much simpler means of making type-2 problems tractable.

The type-1/type-2 distinction may be thought of as a generalization of the distinction between linearly separable and linearly nonseparable problems, according to C&T. This is because type-2 problems require finding relational properties in the input data. It has been known for a long time that problems of this kind present significant difficulties for networks of processing units with monotonic activation functions (see for example Minsky & Papert 1988). C&T consider only learning systems which use processors with monotonic activation functions. They do not consider whether processing units with nonmonotonic activation functions may be able to solve type-2 problems. There are, however, good grounds for believing that networks of such processing units may be able to perform type-2 learning tasks.

Dawson and Schopflocher (1992) describe a kind of processing unit they call "value units"; these have a nonmonotonic Gaussian activation function. Value units can solve linearly nonseparable problems without requiring a layer of hidden units. Recently, Berkeley et al. (1995) have shown how networks of value units can be subject to detailed analysis. In describing their analytic technique, Berkeley et al. discuss a network that became sensitive to a number of relational properties found in a training set of logic problems. These results indicate that value units may have the properties necessary to solve type-2 problems.

To test this hypothesis, a number of networks of value units were trained on the 4-bit parity problem (Rumelhart et al. 1986),

which C&T (sect. 2, para. 17) also studied. All the networks had four input units, a single hidden unit and a single output unit. Sixteen training sets were developed, each consisting of 15 training patterns. The remaining untrained pattern was then used to assess the generality of the solutions discovered by the networks upon convergence. Each training set was trained 11 times, giving a total of 176 simulations. The convergence criterion for each of these simulations was set at 0.025 and in each case a learning rate of 0.01 and a momentum of 0.0 were used. A total of 97 (55%) of these networks reached convergence.

The mean sum squared error for these 97 networks for the seen items in the incomplete training sets was 0.008 and the mean error for the remaining unseen inputs was 0.368. However, such gross figures do not give an entirely accurate picture of the results of training. In contrast to C&T's (sect. 2, para. 18) results in which "In every run, the output associated with the single test item was incorrect," 61 (63%) of the value unit networks generalized correctly when presented with the untrained pattern. The mean error for the unseen patterns for these 61 networks was 0.017 and the mean sum squared error for the training patterns was 0.009. The mean error for the 36 (37%) networks which failed to produce the correct response to the untrained pattern was 0.963, with a mean sum squared error on the training patterns of 0.007. In other words, the results of these simulations show that networks of value units are reasonably successful on "pure" type-2 problems.

These simulation results seem to suggest that another strategy for handling the difficulties presented by type-2 problems is to use a nonmonotonic kind of processing unit. Moreover, this strategy does not involve a commitment to nativism, nor is it task specific, as value unit networks can be trained upon a wide range of problems.

In fact, there is some evidence which suggests that the results from value unit networks may serve to undercut C&T's type-1/type-2 distinction. If C&T (sect. 2, para. 12 and sect. 4, para. 9) are entirely right about the distinction, then the results of the above simulations would be "serendipitous." However, the results of running 25 simulations of the "oddeven" problem described by C&T (sect. 2, para. 9) and illustrated in their Tables 1 and 2 suggests that this is not the case. Training runs on this problem used a value unit network with two input units, a single hidden unit and a single output unit. All networks which reached convergence (20%) had remarkably similar patterns of activation in the hidden unit for each of the inputs in the training set. In other words, each network which converged discovered substantially the same solution to the problem. This is illustrated in Table 1.

These results seem to suggest that the converged value unit networks solve the problem in *exactly* the same way, each time. More important, it appears as if the solution discovered by the networks in each case captures the indirect justification for the output rule which C&T (sect. 2, para. 11) propose (this can be seen by comparing the value of x4 for each pattern with the mean hidden unit activity). Such results are hard to reconcile with C&T's (sect. 2, para. 12) claim that "the space of indirect justifications is infinitely large. To hit on the right one by brute-force search would indeed be 'serendipitous.'" Either value units just happen to be exceptionally suitable for type-2 problems, or C&T's claims about type-2 problems are incorrect. Whichever is the case, it appears that processing units with nonmonotonic activation functions provide a means, in addition to those discussed by C&T, by which "type-2 tigers" (sect. 5, para. 4) can be tamed.

Constraining solution space to improve generalization

John A. Bullinaria

Centre for Speech and Language, Department of Psychology, Birkbeck College, London WC1E 7HX, United Kingdom. **johnbull@ed.ac.uk**

Abstract: I suggest that the difficulties inherent in discovering the hidden regularities in realistic (type-2) problems can often be resolved by learning algorithms employing simple constraints (such as symmetry and the importance of local information) that are natural from an evolutionary point of view. Neither "heavy-duty nativism" nor "representational recoding" appear to offer totally appropriate descriptions of such natural learning processes.

I agree with the main conclusion drawn by Clark & Thornton (C&T) that successful generalization in the case of realistic mappings often requires something more than the simplest statistical analysis. However, I would like to suggest that the case for representational recoding may be somewhat overstated, and that simple constraints on the solution space are often sufficient on their own to lead to good generalization.

Let us consider again the four-bit parity problem cited by C&T. One can explore the solution space in this case without making unnecessary assumptions concerning the properties of particular learning algorithms by performing a Monte Carlo search for solutions in weight space. The minimal network to solve this problem requires four hidden units (and hence 25 degrees of freedom) so we use that architecture. We choose sets of network weights at random (in the range -16 to $+16$) and check to see whether they solve the four-bit parity problem for 15 of the 16 training patterns in the sense that each output unit activation is to the correct side of 0.5. To find 20 solutions took 11.8 billion attempts. Each solution generalized incorrectly to the missing training pattern, which is what we would expect given that random hyperplanes in input space are likely to cut off the missing pattern with its closest neighbors which all produce the opposite output.

We have to ask why we consider one particular generalization to be better than the others. In the sense of Occam's razor, such as embodied in Bayesian model comparison (e.g., MacKay 1992), the best (or "most correct") generalization is the one provided by the

Table 1 (Berkeley). *Results of training value unit networks on the "oddeven" problem*

Pattern Number	Input pattern (x1)	(x2)	Desired output (y1)	Derived Recoding (x4)	Mean Hidden Unit Activity	Variance in Hidden Unit Activity
1	1	2	1	1	4.904e-01	1.497e-03
2	2	2	0	0	9.999e-01	7.500e-06
3	3	2	1	1	4.939e-01	2.065e-03
4	3	1	0	2	5.869e-02	3.840e-04
5	2	1	1	1	4.921e-02	1.279e-03
6	1	1	0	0	9.999e-01	7.000e-07

simplest model (e.g., the one with the least "free" parameters). In fact, smaller networks are well known to provide superior generalization (e.g., Baum & Haussler 1989). In this respect, the arguments of C&T would have been more convincing if six- or more bit parity were used, so that the mapping could be carried out with fewer free parameters (i.e., weights) than training patterns. Since avoiding local minima in minimal six- (or more) bit parity networks is extremely difficult and since it is unlikely that real brains use minimal networks we shall pass over this point.

One natural way to achieve model simplification is by constraining the search space, and one natural constraint might be the imposition of symmetry, that is, start learning assuming maximal symmetry and only relax that assumption as each level of symmetry is found to fail to exist. This will automatically reduce the effective number of free parameters. For example, imposing a symmetry on the weights is sufficient to give good generalization for the four-bit parity problem. Here we constrain the weight solutions to lie on the hyperplanes in weight space corresponding to weights that are symmetric with respect to the input units. This might be implemented in a learning network by constraining the weight changes to be the same for each input unit. This reduces the problem to 13 degrees of freedom and requires only 16.3 million random attempts to find 20 solutions. The symmetry guarantees that all these solutions will generalize correctly. Such "weight sharing" is known to improve generalization more generally (e.g., Nowlan & Hinton 1992).

Another natural constraint we may impose is to assume that local information is more important than distant information until such an assumption is proven incorrect. We may view this to be at work in Elman's grammar acquisition network as discussed by C&T. Elman (1993) implemented these constraints with incremental learning schemes. This is in fact another poor example, since the network not only fails to generalize but also has insufficient processing power to learn even the raw training data (Elman 1993, p. 76). A more powerful recurrent network, or a network with appropriate input buffers or time delay lines, should not have this problem, but there is no reason to suppose that this would improve generalization as well. In time-buffered networks we can constrain solutions to make maximal use of local information by having a smaller learning rates for weights corresponding to longer range dependencies. This approach has also, for example, been shown to improve generalization in past tense acquisition models for which the inflection is usually, but not always, determined by the final phoneme of the stem and in models of reading aloud for which long range dependencies are relatively rare (Bullinaria 1994). Similar constraints may be implemented by weight decay and are also known to improve generalization (e.g., Krogh & Hertz 1992).

Simple constraints on the weight space may not be sufficient to improve generalization for all type-2 problems, but the examples given above indicate that it does have a range of applicability. One might argue that such constraints are just a convenient way to implement the representational recodings of Clark & Thornton, but if that is the case we would seem to have a continuous spectrum of constraints and their type-1/type-2 distinction becomes rather fuzzy.

What is the type-1/type-2 distinction?

Nick Chater

Department of Psychology, University of Warwick, Coventry, CV4 7AL, United Kingdom. **nick@psy.ox.ac.uk**

Abstract: Clark & Thornton's type-1/-2 distinction is not well-defined. The classes of type-1 and type-2 problems are too broad: many nocomputable functions are type-1 and type-2 learnable. They are also too narrow: trivial functions, such as identity, are neither type-1 nor type-2 learnable. Moreover, the scope of type-1 and type-2 problems appears to be equivalent. Overall, this distinction does not appear useful for machine learning or cognitive science.

1. Why probabilities? Clark & Thornton (C&T) frame the learning problem as deriving a conditional probability distribution $P(Y|X)$, where X and Y are sets of possible inputs and outputs, from a set of input-output pairs, (x, y). This is puzzling, because the learning systems that C&T consider (e.g., feedforward neural networks) produce a *single* output, given each input, rather than a conditional probability distribution over all possible outputs.[1] Moreover, C&T state that if a pattern (x, y) has been encountered, then $P(y|x) = 1$ (sect. 2, para. 4), which indicates that they assume that the conditional probability distribution is degenerate – that is, for each input there is a single output. So they appear not be concerned with learning arbitrary conditional probability distributions, but rather with learning *functions* from input to output.

2. All conditional probability distributions are Type 1 learnable. C&T say a distribution to be learned "$P(y|x) = p$ might be [type-1] justified if . . . $P(y|x') = p$, where x' is some selection of values from input-vector x . . ." (sect. 2, para. 4). Suppose x' is the selection of *all* values of x – that is, $x' = x$. Then it trivially follows that $P(y|x) = p$ if and only if $P(y|x') = p$. That is, all conditional probability distributions, including as a special case all functions (including the uncomputable functions), are type-1 learnable.

Note that adding the stipulation that x' cannot include all of x does not help. This can be circumvented by adding irrelevant "dummy" values to each input vector (e.g., a string of 0s) – the learning problem is now just as hard as before. Then the selection x' does not take all elements of the input; it ignores the dummy values. But as before $P(y|x) = p$ if and only if $P(y|x') = p$.

3. The problem of novel outputs. From the above, it seems that C&T's definition does not capture their intuitive notion successfully. From the examples they give, it seems that they intend that $P(y|x')$ is not an arbitrary probability distribution, but rather that it is estimated from frequencies in the input data by $F(y, x')/F(x')$, where $F(x')$ is the number of occurrences of patterns which match x on the selection x' of values, and $F(y, x')$ is the number of such patterns associated with output y.

But this definition is also inadequate in general, because it means that any novel output y_{novel} must be assigned probability 0, because $F(y_{novel}, x') = 0$, precisely because y_{novel} has not occurred in the training set. This means that the class of type-1 problems is very restrictive. It does not include the *identity* function (in which each input is mapped to a different and hence novel output).

C&T face a dilemma. If they follow their stated definition, then all conditional probability distributions are type-1 learnable. If they follow the frequency-based analysis they use in the text, then no conditional probability distribution which assigns a nonzero probability to any unseen output is Type 1 learnable, which seems drastically restrictive.

Furthermore, the frequency-based approach also faces the problem that probabilities can never be estimated exactly from a finite amount of data, and therefore that the $F(y, x')/F(x')$ will not in general equal $P(y|x') = p$. The best that such an estimate can be is probably approximately correct, in some sense (e.g., Valiant 1984).

4. What does type-2 learning mean? C&T say a distribution to be learned "$P(y|x) = p$ might be [Type 2] justified if . . . $P[y|g (\in X) = z] = p$, where g is some arbitrary function, $\in X$ is any seen input, and z is the value of function g applied to x." (sect. 2, para. 4).

This formulation is difficult to interpret, because it uses notation in an unconventional way. But from the later discussion, the appropriate interpretation appears to be this: the function g maps some subset S of previously seen inputs onto a common output, z. We estimate the conditional probability (presumably that which C&T call "$P[y|g (\in X) = z] = p$") by the number of members of S which produce output y, divided by the total number of members of S.

As with type-1 problems, this means that the conditional proba-

bility of all novel outputs must be zero for a problem to be type-2 learnable, for the same reason: the frequency count for novel outputs is necessarily 0. So the identity function is not type-2 learnable either.

But worse, *all* the *non*novel outputs can be justifiably predicted with probability 1. Suppose that a previous input, x_{prev} was paired with the output y_{prev}. Then define g such that $g(x) = z$ (where x is the novel input), and $g(x_{prev}) = z$; $g(x_{rest}) = z + 1$, for all other previously seen inputs x_{rest}. g is a "recoding" of the inputs that classifies the novel input x with a single past input x_{prev}. The subset, S, defined above, has one member, which produced output y_{prev}, so that the estimated conditional probability is $1/1 = 1$. Hence, the arbitrary output y_{prev} is justifiably predicted with probability 1. An analogous argument extends not just to a single novel x, but to all possible novel x. In short, any function whatever which generalizes from the seen instances to the unseen instances is type-2 learnable, even the noncomputable ones (so long as there are no novel outputs).

Note that type-2 problems appear to have the same (rather bizarre) scope as type-1 problems. They are both too broad and too narrow in the same way.

NOTE

1. The output of neural networks can be viewed as a probability distribution over possible outputs if, for example, outputs are binary and intermediate values are interpreted as probabilities (e.g., Richard & Lippman 1991). A different approach assumes that outputs are distorted (for example by Gaussian noise). This is useful in understanding learning in Bayesian terms (Mackay 1992). Moreover, some networks implicitly produce conditional probability distributions by generating a distribution of outputs over time (e.g., the Boltzmann machine; Hinton & Sejnowski 1986). None of these approaches seems relevant to C&T's discussion.

Parity is not a generalisation problem

R. I. Damper

Cognitive Sciences Centre and Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, England.
rid@ecs.soton.ac.uk; www-isis.ecs.soton.ac.uk

Abstract: Uninformed learning mechanisms will not discover "type-2" regularities in their inputs, except fortuitously. Clark & Thornton argue that error back-propagation only learns the classical parity problem – which is "always pure type-2" – because of restrictive assumptions implicit in the learning algorithm and network employed. Empirical analysis showing that back-propagation fails to generalise on the parity problem is cited to support their position. The reason for failure, however, is that generalisation is simply not a relevant issue. Nothing can be gleaned about back-propagation in particular, or learning in general, from this failure.

1. Introduction. Clark & Thornton (C&T) argue that many learning problems involve "type-2" mappings, characterised by "attenuated existence in . . . training data." Thus, their discovery by an uninformed learning device (such as back-propagation) presents intractable problems of search. Once serendipitously found, however, the type-2 mappings can be exploited in further, modular learning. It is hard to argue against the principle of such recoding playing an important part in learning: indeed, it is almost a truism in cognitive science. Rather, I wish to show that C&T's detailed supporting arguments based on the empirical inability of back-propagation to generalise on the classical parity problem are mistaken.

2. Parity, generalisation, and mind reading. As with many others, my interest in neural computing blossomed when I obtained McClelland and Rumelhart's *Explorations in parallel distributed processing* in 1988. As its advertised purpose was to encourage experimentation, I tried to get the **bp** program to generalise on the 2-variable parity (XOR) problem. Given a network with 2 input nodes, 2 hidden nodes, and a single output, together with the first three lines of the truth table:

x_1	x_2		y
0	0	⇒	0
0	1	⇒	1
1	0	⇒	1

could **bp** generalise to find the missing line:

1 1 ⇒ 0 ?

I soon realised that was a silly thing to expect. How could *any* machine learning procedure generalise on this problem? The y-output for the unseen mappings is entirely arbitrary and hence unpredictable from the three seen mappings, although the unconditional probability $P(y = 1) = 0.67$ on C&T's argument favours a 1 output corresponding to a learned OR solution. Expecting the back-propagation network to generalise consistently to the XOR function – rather than the simpler OR function – *solely* on the basis that this is what in the experimenter's mind, amounts to expecting the network to be a mind-reader. Parity is not a generalisation problem! Yet this seems to be a cornerstone of C&T's thesis. To quote: "parity cases . . . do not really warrant any optimism concerning the chances of backpropagation . . . hitting on the right recodings." Thus, the inability to generalise on the parity problem is taken to have implications for cognition when, clearly, it does not.

This inability is apparently well-known to Elman who writes (1993, p. 85): "If the fourth pattern (for example) is withheld until late in training, the network will typically fail to learn XOR. Instead, it will learn logical OR since this is compatible with the first three patterns." Presumably, by "compatible" he means the unconditional probabilities favour the OR solution. As pointed out above, however, the polarity of the output for the unseen input combination is arbitrary. To this extent, both XOR and OR are equally "compatible" with the first three patterns.

I ran several simulations on the first three patterns using **bp.** In 20 out of 20 repetitions, the linearly-separable OR solution was always found. The hidden-layer representation was always that of one unit having weights and biases corresponding to an OR separating "hyperplane" like that labelled *line 1* on Figure 1 while the other "did nothing," that is, its weights and biases corresponded to a line which did not separate the input patterns at all. While there is a finite chance that this "do nothing" line just

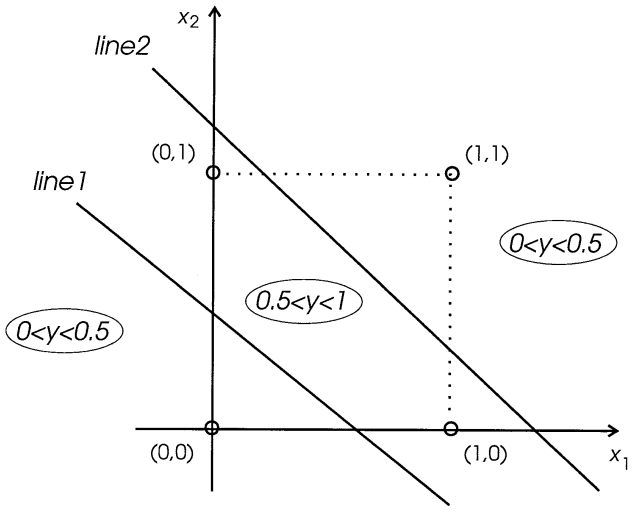

Figure 1 (Damper). Possible separating "hyperplanes" for the 2-input parity problem, corresponding to the decision boundaries formed by the two hidden units. In this simple case, the hyperplanes are lines.

happens to separate the unseen ($x_1 = 1$, $x_2 = 1$) input from the three seen patterns – so learning the XOR function – it is remote and never occurred in my simulations. In no way do these results reflect the unconditional probability $P(y = 0) = 0.33$ – although back-propagation can be used to estimate such probabilities with an appropriate cost function replacing the squared error measure (Baum & Wilczek 1988). The OR solution was found very fast: typically in 55–65 iterations with learning rate 0.5, momentum 0.9, and error criterion of 0.03 (based on an average allowable error of 0.1 per pattern).

3. "Extended" parity problem. Although parity itself – involving binary-valued variables – is not a generalisation problem, it is straightforward to extend the situation depicted in Figure 1 to a *bona fide* generalisation problem. Suppose our training data-set is:

x_1	x_2		y
0	0	⇒	0
0	1	⇒	1
1	0	⇒	1
1.5	0	⇒	0
0	1.5	⇒	0

This is no longer a parity problem, since the variables are no longer binary but continuous. However, the last two additional data points are now sufficient for back-propagation to find a hidden-layer representation which places *line 2* "correctly" (e.g., as in Fig. 1, although other solutions to the parity problem – and this "extended" version – are possible). In 10 repeated trials of **bp** with values of learning rate and momentum set to 0.2 and 0.9 respectively (as used by C&T), convergence to an error criterion of 0.04 was obtained within 5000 epochs on 8 occasions (mean 1277 epochs). (This compares to corresponding figures of 9 successful trials out of 10, mean 1366 epochs, for the parity problem trained on all four entries of the truth table – which is not significantly different. In fact, rather faster training with a mean of 337 epochs to reach the error criterion was obtained in both cases using a learning rate of 0.5, which is McClelland and Rumelhart's default value.) In all 8 cases, the unseen input ($x_1 = 1$, $x_2 = 1$) was classified perfectly as binary 0, with the output activation in the range 0.01–0.04.

4. Discussion. What are the implications of this for C&T's thesis? First, nothing can be gleaned about either machine or animal learning from the fact that back-propagation fails to generalise on the parity problem. Generalisation is plainly not relevant to the situation. Generalisation is only possible when there is regularity within classes of similar input patterns, but parity "is a very hard problem, because the output must change whenever any single input changes" (Hertz et al. 1991, p. 131).

Second, as stated by C&T themselves, back-propagation *does* discover the solution to parity problems relatively easily, in spite of this being "a very hard problem." They attribute this to restrictive assumptions – the network must have appropriate structure, parameters, and so on. (It must also be shown all the data and not be expected to mind-read!) Yet back-propagation is actually one of the least promising learning techniques as far as discovering recodings is concerned. Its popularity in cognitive science derives partly from Sutton and Barto's (1981) observation that the delta rule (of which back-propagation is a generalisation) is near identical in form to Rescorla and Wagner's (1972) model of classical conditioning, thereby lending iterative gradient-descent search some psychological respectability. However, non-iterative, constructive techniques have much less difficulty in learning parity – for example, as an and-or network, by simple observation of the $(0, 1) \Rightarrow 1$ and $(1, 0) \Rightarrow 1$ entries in the truth table – and are at least as plausible.

Finally, I think it is necessary to be cautious in using "toy" problems (like parity) to reason about complex, real-world situations. Certainly, toy problems can bring to the fore important issues which might otherwise be obscure, but they are characterised by exact solutions, binary variables, etc. – quite unlike most real-world problems.

Epistemological missing links

Terry Dartnall

Computing and Information Technology, Griffith University, Brisbane, Australia 4116. **terryd@cit.gu.edu.au**

Abstract: Clark & Thornton's "superficially distinct ploys and mechanisms" are in fact very different: there is a deep difference between (a) filters and feature detectors, which "let the information in," and (b) contentful representations and theories, which reconfigure it into a computationally tractable form. (a) is bringing *abilities* to experience whereas (b) is bringing *content* to experience. Both have well known problems. I outline an evolutionary story that avoids these problems and begins to explain how representations and theories developed out of feature detectors and filters.

Clark & Thornton (C&T) talk about "a wide variety of *superficially distinct* ploys and mechanisms . . . [that] range from (a) simple evolved filters and feature-detectors all the way to (b) complex cases involving the use and re-use of acquired knowledge" (my emphasis). These superficially distinct mechanisms are really deeply different: (a) is the ability to focus on relevant features of the environment using filters and feature detectors. This gives us direct access to the data, but focuses on certain aspects of it, or feeds bits of it to us a little at a time; (b) is the ability to bring theories (or models, maps, representations, ideas – in general, *content*) to experience, to "systematically reconfigure it" into a computationally tractable form that can be tackled by what C&T call "uninformed search."

The distinction between bringing abilities and content to experience has been a major issue in Western epistemology. Rationalists say we bring content; empiricists say we bring only abilities (association, combination, search, etc.). Both approaches have problems. If we bring content, then in some respects the world of experience is the product of our conceptualisation, and hence does not exist independently of us. According to classical empiricism, sensory impressions give rise to representations, and these are what we experience. Thus the world is either partly the product of our conceptualisation, or is hidden behind a representational veil.

I suggest that the answer lies in something that C&T discuss: "direction of fit." According to C&T, a problem in Elman's (1993) work on grammar acquisition is that "it is not immediately clear why the veil of restricted short-term memory should filter the [linguistic] data in just the right way." They suggest that the answer lies in Newport's (1990) theory that the child's early capacities were selected to facilitate language learning. There is such a nice fit between the structure of language and the child's limited early capacities to handle complex stimuli because the structure of language was selected to exploit the limited nature of early short-term memory in human infants.

That is okay for language; but what about direction of fit in a more general sense? We experience appropriate features of the world, and learn to do so in an appropriate developmental sequence. We can think about the world, and to some extent understand and predict it. Is this because our cognition contributes to the structure of the world, or because nature has selected mechanisms for understanding her? The first possibility forces us to say that we can only experience and think about the world of experience, not about the world as it really is. The other possibility enables us to say what we surely want to say: that we directly experience a world that exists independently of us. It also gives us important insights. The cognitive capabilities of our earliest ancestors evolved to cope with features of the world that were necessary for survival. More sophisticated filters and feature detectors developed as competition became more demanding, and early mechanisms were adapted to this end. Let us assume that the original mechanisms were left in place, so that the individual inherited a layer of increasingly sophisticated mechanisms triggered sequentially during early development. And let us assume that, once triggered, they could be drawn on by the individual at will.

This evolutionary story makes no mention of representations. In fact, it can account for some of the things that representations have

been invoked to explain. In Karmiloff-Smith's (1992) Representational Redescription Hypothesis, a layer of representations can be accessed at will [see multiple book review of "Beyond Modularity" *BBS* 17(4) 1994]. In our account, a layer of *mechanisms* can be accessed at will. A standard example of representational redescription is when data stored as a stack are redescribed as a list. (Items in a stack can only be accessed from the top, but items in a list can be accessed in any order.) Suppose that at some point in our evolution we could only access items or perform actions in a stack-like sequence, and later we learned to access or perform them in any sequence. Now we *could* say that we learned to represent the world differently, and that this in turn gave us better access to it, but it is simpler to say that we learned to access it more efficiently. This is a story about mechanisms, not representations.

The story sheds light on abstraction. If by "abstraction" we mean "getting at the general features," then a system of layered mechanisms would give us this ability. Early in our evolution we focussed on the most salient features of the environment (edges and corners, high and low pitches of sound, etc.), but we gradually developed mechanisms for accessing it in more detail. We accordingly need only revert to our earlier mechanisms to pull out the most general features of the environment. Rather than doing a search for general features, we revert to our lowest-level detectors and filters. So abstraction, our most prized intellectual possession, may not be what it seems.

This story avoids the problem of the representational veil. But we can still have representations and maps. The condition is that we must be able to "get behind them," to check up on the information they are trying to organise. We might bring multiple maps to experience, but we need fast ways of checking them out. The driver who only looks at the map is not an evolutionary success. Thus, on the one hand we have mechanisms that "let the information in," and on the other we have manipulable, controllable structures that enable us to reduce the complexity of subsequent search. And we have a sequence: filters and feature detectors first, maps and representations later. It is reasonable to assume that the latter exploited the former. But how?

We know that filters and feature detectors are more efficient if they are accompanied by domain knowledge: the more we know about the domain, the less search we have to do. For example, it is easier to understand a sentence from which letters have been deleted if we know that the sentence is a proverb. Representations and maps also depend on accompanying knowledge, more so as the domain becomes more abstract. This ability to shift the cognitive load from specific filters and feature detectors to general knowledge about the domain must have been a major step forward. So was the ability to have accompanying knowledge, not about the domain ("this is a proverb"), but *about the filter or feature detector itself* ("this only picks out proverbs . . . or corners . . . or edges"). Such knowledge enables us (a) to choose between filters within a domain and (b) to use them across domains, thus overcoming domain-specific constraints. This is a tall order, but it brings us closer to C&T's multiple maps and representations.

Reducing problem complexity by analogical transfer

Peter F. Dominey

INSERM U94 69500 Bron; CNRS EP-100 69008 Lyon, France.
dominey@lyon151.inserm.fr

Abstract: Analogical transfer in sequence learning is presented as an example of how the type-2 problem of learning an unbounded number of isomorphic sequences is reduced to the type-1 problem of learning a small finite set of sequences. The commentary illustrates how the difficult problem of appropriate analogical filter creation and selection is addressed while avoiding the trap of strong nativism, and it provides theoretical and experimental evidence for the existence of dissociable mechanisms for type-1 learning and type-2 recoding.

Clark & Thornton (C&T) cite analogical reasoning as an example of how previously learned concepts can be used to filter out the salient regularities in novel situations in order to reduce type-2 problems to type-1 status. This commentary addresses the important open issue of how such filters might be developed and chosen while avoiding the trap of strong nativism. The trap is that one might require a specific mechanism for representing each different concept/filter, and thus be no better off than without analogical reasoning. What would be more appropriate is a single mechanism that could provide a general capacity for analogical reasoning.

In this context, it has been proposed that the recognition of structural isomorphisms is one of the primary means by which parts of two analogs can be placed in correspondence with each other (Thagard et al. 1990). For example, consider the two sequences, ABCBAC and DEFEDF. While they share no common surface structure, these isomorphic sequences share the abstract relational structure "*u*, *u*, *u*, *n*-2, *n*-4, *n*-3," where *u* indicates unique or nonrepeated (unpredictable), and *n*-2 indicates a repetition predictable of the element 2 places behind, and so on. The ability to store, recognize, and use this kind of structural isomorphism should contribute to a general mechanism for analogical reasoning in a profitable tradeoff between nativistic predefined functions and robust generalized behavior.

In order to study such a mechanism we developed a test of analogical transfer in sequence learning (ATSL). The test is based on the serial reaction time (SRT) task, in which learning is demonstrated by a reduction in reaction times for stimuli that appear in a repeating sequence versus stimuli that appear in a random series (Nissen & Bullemer 1987). Sequence learning can occur in uninformed or implicit conditions, that is, the statistical regularities in the sequence can be extracted by an uninformed type-1 mechanism. In the ATSL task, however, the same sequence is never repeated. Instead, a number of isomorphic sequences are successively presented. This is a type-2 problem in that the statistical regularities of the potentially unbounded number of sequences become visible only when they are recoded in terms of their shared relational structure.

We have recently observed that normal human subjects are capable of this kind of recoding in the ATSL task, that is, they display learning in the form of monotonically decreasing reaction times for predictable stimuli in a series of isomorphic sequences (Dominey et al. 1995b). It is interesting to note that this type-2 learning is only observed in subjects who have been explicitly informed that such an abstract structure might exist. Implicit or noninformed subjects display no such learning, in striking contrast with their known capacity for type-1 learning in the SRT task (Nissen & Bullemer 1987).

To confirm the observation that this type-2 task cannot be performed by a type-1 system, we performed simulation studies using a model of type-1 sequence learning based on the neural architecture of the primate frontostriatal system (Dominey et al. 1995a; Dominey 1995). In the model, learning-related synaptic modifications generate increased activation of appropriate response units for stimuli in learned sequences, with a corresponding RT reduction. Due to this property, the model demonstrates type-1 SRT sequence learning when using a single repeating sequence (Dominey 1996). It fails, however, in the type-2 ATSL task, with the same lack of learning as observed in the implicit learning group (Dominey 1996; Dominey et al. 1995b).

For the model to exploit the abstract structure shared by isomorphic sequences like ABCBAC and DEFEDF, it must be capable of representing such sequences in terms of the structural relation sequence "*u*, *u*, *u*, *n*-2, *n*-4, *n*-2" that is common to them. This requires (1) a form of short term memory (STM) of the several previous elements in a sequence, and (2) the capacity to recognize whether the current element matches any of those in the STM, for example, that the second A in ABCBAC matches the element 4 places behind. Finally, this recoding of the sequences must be made available to the existing type-1 learning mechanism.

The type-1 sequence learning system can then be used to learn this abstract structure that serves as a filter for subsequent inputs. In the same way the type-1 system alone can predict specific elements in learned sequences, the modified type-2 system can predict repeated elements in learned classes of isomorphic sequences. Indeed, we observed that the modified type-2 model reproduces the performance of explicit subjects in the ATSL task (Dominey et al. 1995b). The type-2 mechanism is simultaneously capable of (1) applying "filters" learned from previous experience, that is, recognizing learned abstract structures in order to predict the repetitive structures of new isomorphic sequences, and (2) developing new "filters," thus learning new abstract structures for isomorphic sequences whose abstract structure has not previously been learned. The system achieves this by continuously exploring, in parallel, the space of possible abstract structures, recognizing learned structures and learning new ones as necessary. The filters (abstract structures), are stored as remembered sequences, and are selected by a type-1 process of sequence recognition. Note that this type-2 mechanism should generalize to the related problems of (1) exploiting several abstract structures in a body of input data, and (2) grammaticality judgment after letter set transfer in artificial grammar learning (Gomez & Schaveneveldt 1994).

From a neurophysiological perspective, it is of interest that type-1 SRT learning is impaired in Parkinson's disease (Jackson et al. 1995), indicating that the frontostriatal system may participate in this type-1 learning. In contrast, our recent studies of analogical transfer in Parkinson's patients have demonstrated that the impairment in type-1 SRT learning is not seen in type-2 ATSL learning in these patients (Dominey et al. 1996). This suggests a functional distinction in the role of the frontostriatal system in type-1 and type-2 learning.

In conclusion, analogical transfer in sequence learning is presented as an example of how the type-2 problem of learning an unbounded number of isomorphic sequences is reduced to the type-1 problem of learning a single or greatly reduced set of sequences, by continuously recoding sequences in terms of their relational structure. This example is of theoretical interest in that (1) it provides an explicit demonstration of how the potentially difficult problem of appropriate analogical filter creation and selection is addressed while avoiding the trap of strong nativism, and (2) it provides theoretical and experimental evidence for the existence of dissociable mechanisms for type-1 learning and type-2 re-coding.

ACKNOWLEDGMENT

The author is supported by the Fyssen Foundation (Paris).

Cognitive success and exam preparation

Matthew Elton

Department of Philosophy, University of Stirling, Stirling, Scotland, FK9 4LA.
matthew.elton@stirling.ac.uk
www.stir.ac.uk/philosophy/staff/elton

Abstract: Evolution is not like an exam in which pre-set problems need to be solved. Failing to recognise this point, Clark & Thornton misconstrue the type of explanation called for in species learning although, clearly, species that can trade spaces have more chances to discover novel beneficial behaviours. On the other hand, the trading spaces strategy might help to explain lifetime learning successes.

Clark & Thornton's (C&T's) target article is about the general principles of operation of cognitive mechanisms. More precisely, they are interested in operational principles that are likely (rather than those that necessarily must or, as it happens, are) to be found in mechanisms that have evolved through natural selection. Hence their paper is about cognitive science, or even evolutionary cognitive science, rather than, say, simply about artificial intelligence.

The principle of operation at issue is learning and at the heart of C&T's paper is a distinction between two sorts of learning problem. Statistical approaches to learning will find type-1 problems fairly easy, C&T explain, and, in the absence of suitable re-coding, type-2 problems very hard. One thing that is attractive about this central claim is the level at which it is pitched. Rather than simply undertaking empirical trials, comparing this connectionist system with that symbolic one, or that network architecture with another, they have also sought to identify a general principle that operates independently of a connectionist-symbolist dispute. While connectionist learning systems do have many advantages, they can no more work magic with a type-2 problem than any other form of statistical learning. This sort of thesis is much needed in cognitive science. If it holds up, then, at a stroke, it can transform empirical hunches about the power of this or that algorithm into firm theoretical results.

C&T describe an animat that starts life lacking the skill of "conditional approach." Its problem is to acquire this skill, but, because the acquisition problem is type-2, the animat is unlikely to stumble across the solution without trading spaces, even if given a great deal of help in the form of training. (And, if I understand C&T aright, it is given rather more help than an animal in a natural setting could expect.)

Although an effective illustration of the type-1/type-2 distinction, the example uncovers a curious inversion in C&T's thinking. Imagine the evolution by natural selection of a creature such as their animat – call it a "clarkton." At some stage, the clarkton is in the same state as the unschooled animat. Let us suppose the environment is putting it under increasing pressure and that the clarkton would greatly benefit from acquiring "conditional approach." Consider three possible futures: (i) the clarkton, by very lucky chance, or with less lack and the application of the trading spaces strategy, solves C&T's problem, acquires "conditional approach," and so improves its fitness; (ii) the clarkton solves another quite different type-2 problem, one which improves its fitness in such a way as to obviate the benefit of "conditional approach"; (iii) the clarkton fails to improve its fitness and dies. (Note that if we forget the high frequency of outcomes like [iii] we gain a false impression of the efficiency of natural selection.)

The only difference between (i) and (ii) is that in (i) the behavioural change is one in which C&T have a special interest. It is only with hindsight that the clarkton can be seen to have "solved" C&T's "problem," and this shows that talking about problems and searches for solutions is out of place here. Evolution is not like an exam, where the problems are set ahead of time. Rather, many different routes are tried out, and creatures stick with those that work. Creatures don't "aim" to acquire specific skills, though when they do acquire a new skill and survive, it is usually a skill worth having.

By contrast, lifetime learning can be rather like an exam. A creature has to keep tackling just one problem until it gets it right. Here the trading spaces idea might make a real contribution, explaining how a strategy of repeatedly recoding and trying again (or perhaps trying different codings in parallel) improves one's chances. Of course, chance is still involved in stumbling across the right kind of coding.

But not only chance. C&T argue that the sorts of problems that, say, human beings have become adept at solving are structured in such a way as to be amenable to our problem solving powers. Lifetime learning is like an exam, but the exam script consists of problems at which earlier candidates were successful. This analogy is too crude: the development of language, say, must surely have involved very many incremental cycles, with small dividends at each stage, in which complexity was slowly ratcheted up. But the route of development was such that each individual was well equipped to "solve" the relevant "learning problem" at each stage. To the extent, however, that this part of the argument is a success, it begins to lessen, though not eliminate, the need to explain type-2 problem solving powers.

By thinking of evolution as being like an exam, C&T create a spurious difficulty. It is only with hindsight that natural selection "solves" its design "problems." Any successful evolution of a

behavioural strategy can, after the event, be seen as the solution to some problem, whether of type-1 or type-2. This said, a species that is capable of trading spaces is able to stumble across a richer range of new behaviours, so I don't dispute that the strategy could play a part in evolutionary explanations. To the extent that lifetime learning problems are externally set, and are type-2, the trading spaces strategy does offer an insight into how they might be solved. Learning algorithms aren't the whole story, however and, the authors do steer a subtle course between nativism and empiricism. Their claim effectively is that problems on individuals' life exams are ones for which, for historical reasons, their cognitive mechanisms are particularly well-prepared. And so it turns out that in both lifetime and species learning our problems look more difficult than they really are.

Type-2 problems are difficult to learn, but generalize well (in general)

M. Gareth Gaskell

Centre for Speech and Language, Department of Psychology, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom.
g.gaskell@psyc.bbk.ac.uk

Abstract: Learning a mapping involves finding regularities in a training set and generalization to novel patterns. Clark & Thornton's type distinction has been discussed in terms of generalization, but has limited value in this respect. However, in terms of detection of regularities in the training set, the distinction is more valid, as it provides a measure of complexity and correlates with the size of search space.

We can ask two basic questions about the performance of a connectionist network when applied to an input-output mapping. First, does the network perform well on the training set? Second, does it perform well on novel input-output pairs selected from the same population as the training examples? In other words, has it learnt and will it generalize?

The type-1/type-2 distinction appears to be set up to address the second question, and the examples Clark & Thornton (C&T) provide are cited as failures to generalize from correctly learnt training examples. Certainly, for the 4-bit parity example, the networks performed at near ceiling level on the training examples but failed to generalize to the held-back pattern. Part of the problem here is that the number of training examples is small and the number of free parameters is relatively large, implying that there are many ways of solving the problem for the training set, many of which will not be applicable to the unseen pattern.

However, the problem of choosing between multiple suitable recodings, given a small number of training examples, is not restricted to type-2 mappings. Consider the training set shown in Table 1. Here, all conditional probabilities are extreme, and multiple type-1 solutions can be found. However, there is no guarantee that the one chosen will agree with the underlying regularity in the set from which these mappings were selected. For example, in the overall set, $x3$ might be the only reliable predictor of the value of $y1$, but a standard network is likely to pay more attention to the values of $x1$ or $x2$. Thus, when the number of training patterns is small, multiple solutions may exist for both type-1 and type-2 problems. Consequently, generalization to novel items cannot be guaranteed for either type of mapping.

The above example is of course artificial, since most tasks faced by the human system have a much larger scale. In these cases, the real difficulty lies in learning to perform the correct mapping on the training examples. Once a solution is found that is applicable to the large (and hopefully representative) set of training examples, the chances of generalization to novel patterns are good. The type-1/type-2 distinction is no doubt valid as a broad classification of complexity in this case. As C&T point out, type-2 problems involve relational learning, which requires transformation of the input. This increases the number of free parameters and thus the size of the solution space.

Table 1 (Gaskell). *Input-output mappings for a type-1 problem with multiple solutions*

$x1$	$x2$	$x3$	$x4$	$y1$
1	0	0.51	0.25	1
0	1	0.49	0.75	0

This point is nicely exemplified by Elman's (1993) simulations, which C&T cite as an example of a simple failure to generalize (sect. 3, para. 15). In fact, the network does not perform satisfactorily on even the training set unless either the network or the training set is altered to obscure long distance dependencies (Elman 1993, p. 76). Instead, the problems the network encounters are local minima in the training error curve, which are avoided by ensuring that initial learning moves the network to a region in weight space from which the global minimum can be reached. Once the network has found a solution that is applicable to the large, complex, and representative training set, generalization is more or less guaranteed.

Thus, the type-1/type-2 distinction captures one aspect of the problems involved in learning mappings. For type-2 mappings, the search space is comparatively large and successful performance on the training data may require additional constraints that make the exploration of this space more tractable. However, the distinction is less applicable to the generalization problem, which simply requires a good balance between the number of variables in the mapping function and the number of data points that the learning device is exposed to. Generalization will fail when there are too few data points or too many free parameters, leading to a plethora of possible solutions with no way of choosing between them.

Model-based learning problem taxonomies

Richard M. Golden

School of Human Development, University of Texas at Dallas, GR 41, Richardson, TX 75083-0688. **golden@utdallas.edu; www.utdallas.edu/~golden**

Abstract: A fundamental problem with the Clark & Thornton definition of a type-1 problem (requirement 2) is identified. An alternative classical statistical formulation is proposed where a type-1 (learnable) problem corresponds to the case where the learning machine is capable of representing its statistical environment.

An important feature of the Clark-Thornton definition. Note that requirement (2) in the Clark & Thornton (C&T) definition of a type-1 learning problem implies that the problem where $x_1 = [0, 0, 0]$, $x_2 = [0, 0, 1]$, $x_3 = [0, 1, 0]$, $x_4 = [1, 0, 0]$, $x_5 = [1, 1, 0]$, and $y_1 = 0$, $y_2 = 0$, $y_3 = 1$, $y_4 = 1$, $y_5 = 1$, and one must discover that $y_i = x_{i1}x_{i2}$ (where x_{i1} and x_{i2} are the first two elements of x_i) is a type-1 learning problem. This type of problem, for example, is relatively easy for a linear perceptron-like learning rule to solve.

Now consider a modification of the C&T definition of a type-1 learning problem designed to emphasize the importance of requirement (2) of the definition. If requirement (2) was modified to read "$P(y|x')$ where $x' = x$," then the above learning problem (which can be solved by linear perceptron-like learning rules would be defined as a type-2 learning problem, which would be an undesirable characteristic of the definition. Thus, requirement (2) as formulated by C&T is quite appropriate.

A problem with the Clark-Thornton definition of a type-1 problem. The problem with C&T's requirement (2) is that the number of subsets of the d-dimensional vector x which must be examined

is 2^d which is an exponential function of the dimensionality d of x. Thus, searching for the "right subset" of x is equivalent to searching for the "right mapping function" from a set of 2^d possible functions (which map x into a subset of x).

A classical model-based definition of type-1 learning problems. Rather than trying to "abstractly" classify data sets independently of models as either: (i) type-1 (easy), or (ii) type-2 (hard) problems, a "parametric model-based" approach is proposed. This alternative explicitly defines a probability model (i.e., a set of probability distributions) for a given learning machine that specifies which probability distributions the learning machine is capable of implicitly representing. Define a type-1 (learnable) statistical environment (i.e., the probability distribution of the data generating process) with respect to some specific probability model $\mathcal{M}$ as a probability distribution which is an element of $\mathcal{M}$. A type-2 (unlearnable) statistical environment with respect to $\mathcal{M}$ is a probability distribution which is not an element of $\mathcal{M}$. Statistical tests can be constructed (i.e., goodness-of-fit tests) to infer whether a given statistical environment is type-1 or type-2 with respect to a learning machine's given probability model. Golden (forthcoming, Ch. 7, Ch. 8; also see Golden 1988 and White 1989) has explicitly shown how this type of classical statistical framework is applicable to a wide variety of connectionist learning problems.

Trading spaces: A promissory note to solve relational mapping problems

Karl Haberlandt

Department of Psychology, Trinity College, Hartford, CT 06106.
karl.haberlandt@trincoll.edu

Abstract: Clark & Thornton (C&T) have demonstrated the paradox between the opacity of the transformations that underlie relational mappings and the ease with which people learn such mappings. However, C&T's trading-spaces proposal resolves the paradox only in the broadest outline. The general-purpose algorithm promised by C&T remains to be developed. The strategy of doing so is to analyze and formulate computational mechanisms for known cases of recoding.

Clark & Thornton's (C&T) target article distinguishes between direct and relational input/output mappings (type-1 versus type-2 problems). In the former, the relation between input and output patterns can be discerned through direct inspection of the input/output pairs themselves. In indirect mappings, input and output pairs are related through any one of an infinite number of transformation functions that are not transparent. Parity problems such as the one in Table 3 (sect. 2) represent cases of indirect mapping (Rumelhart et al. 1986). An important issue for cognitive scientists has concerned how learners discover the hidden relational functions that transform input patterns into the corresponding output patterns.

Backpropagation learning, "currently the most important and most widely used algorithm for connectionist learning" (Gallant 1994, p. 211), has been touted as the algorithm to do just that. Indeed, backpropagation is the discovery that made possible the resurrection of the neural network enterprise. It is therefore newsworthy that C&T have demonstrated limits of backpropagation in acquiring parity-type problems in sections 2 and 3. Significantly, C&T show that backpropagation is limited even when one takes the greatest liberties in designing the network architecture and choosing learning parameters (Haberlandt 1990).

Using backpropagation, C&T sought to teach a 3-layer network a relatively simple 4-bit parity problem. Whereas the network successfully acquired the training cases, it failed to generalize to critical test cases not presented during training. If nothing else, this failure shows that backpropagation does not always work unless the programmer somehow divines the "correct" architecture, parameters, and schedules.

Acquiring the conditional-approach behavior of the animat in Figure 3 (sect. 3) proved similarly difficult under several regimens of backpropagation and other learning algorithms. This problem was so hard to learn because the robot must detect a relation, the "ratio between apparent closeness and apparent width" of the object placed in front of it. When C&T hand-coded the critical relation explicitly, however, they readily succeeded in training the robot.

It is at this point in C&T's article that readers may have their hopes up for more than the promissory note that relational problems of the animat type and others are tractable by trading spaces, which means by "using achieved representation to reduce the complexity of computation." What is needed is a computational mechanism that detects the relations that transform input into output patterns without programmer intervention or assistance. I worry that if such a mechanism cannot be formulated for the relatively straightforward animat problem it is less likely to be discovered for the authors' second case study, Elman's (1993) rule learning case.

Elman developed a network model to simulate the acquisition of long-distance dependencies in English sentences involving verb agreement and clause embedding. His three-layer network was able to learn sentences with cross-clausal dependencies only in an incremental learning scheme in which simple examples were presented in a first stage and more complex ones in the second stage. The first stage was necessary to fine-tune the network toward the relevant dimensions of the stimuli. C&T's final illustration, Karmiloff-Smith's (1992) notion of representational redescription in cognitive development, may be more abstract than the previous examples. According to C&T, however, it exemplifies the same principle, namely, that achieved representations are traded for complexity in computation.

The concept of "trading spaces" is tantalizing even if no computational mechanism is available, let alone the "general strategies" promised in the abstract of the target article. I find the concept tantalizing because it may be the spark for the development of an alternative to the idea that nature endows us with the knowledge to detect such relations automatically.

A fruitful strategy would be to start by examining conceptual issues using cases of recoding proposed by psychologists. Issues include the following: Which instances of recoding reflect direct and indirect mapping? Does it matter whether the learner is acquiring facts or skills, and whether or not feedback is provided? In which ways do mappings differ across domains? What justifies C&T's optimism for "general strategies" in light of research where expertise is domain specific? There is a plethora of cases of recoding to select from, beginning with Miller's (1956) classical paper on chunking and Reber's (1993) synthetic grammar to Anderson's (1983) proposal on compiling of procedures. Psycholinguistics has been a particularly fertile ground for recoding schemes, including Elman's (1993) work. Recoding is postulated at every level of language processing: from phonemes and graphemes to lexical items (Sejnowski & Rosenberg 1986), from phrases to propositions (Haberlandt et al. 1986; Jarvella 1979), and from propositions to text bases (Kintsch 1988).

The hunt for general algorithms to solve relational mapping problems is on. As Anderson (1995) observed, this is a pressing unsolved problem. He also diagnosed quite accurately that, while simple memory based associators may provide correct input-output pairings for specific cases of mapping, general techniques for reliable generalizations remain to be discovered by future research.

Recoding can lead to inaccessible structures, but avoids capacity limitations

Graeme S. Halford

Psychology Department, University of Queensland, 4072 Queensland, Australia. **gsh@psy.uq.oz.au**

Abstract: The distinction between uninformed learning (type-1) and learning based on recoding using prior information (type-2) helps to clarify some long-standing psychological problems, including misunderstanding of mathematics by children, the need for active construction of concepts in cognitive development, and the difficulty of configural learning tasks. However, an alternative to recoding some type-2 tasks is to represent the input as separate dimensions, which are processed jointly. This preserves the original structure, but is subject to processing capacity limitations.

Clark & Thornton's (C&T's) distinction between type-1 and type-2 learning helps clarify some long-standing psychological problems. Type-2 learning is a major source of difficulty in children's learning of mathematics, because their experience with arithmetical operations provides insufficient constraint for acquisition of the relevant concepts, leading to "malrules" that superficially resemble, but do not conform to, correct mathematical procedures. The intractability of the task of acquiring the knowledge children require for cognitive development has led to the view that children must actively formulate the concepts they acquire. In this respect, constructivism could be seen as a special case of the theory that type-2 learning requires active recoding. Our experimental studies of induction of relational schemas have produced large variance because of the way participants code the task. We realized after a while that it would be a mistake to regard this as simply error-variance, because it is really an important part of our findings.

Conditional discrimination, in which a discrimination between (say) two shapes is reversed on change of background, is type-2, and resembles the structure in C&T's Table 1. Consistent with type-2 theory, this task has always proved difficult, for both animals and children (Halford 1993; Rudy 1991). The standard explanation is that these tasks are learned by forming configurations, or unique combinations of cue and background, which is a form of recoding.

Recoding the input permits the output function to be computed and exemplifies a psychologically realistic mechanism, but can make the structure inaccessible. Notice that the structure in C&T's Table 1 cannot be recreated from C&T Table 2 (the mappings of $x1$ and $x2$ into $y1$ cannot be reconstructed from the mapping of $x4$ into $y1$). In general the original structure is not necessarily recoverable from the recoded structure.

However, there is an alternative which does not have this drawback. This can be illustrated with the training set in the C&T Table 1. While the task is insufficiently constrained by contingencies between either $x1$ or $x2$ and $y1$, it is adequately constrained by $x1$ and $x2$ taken jointly. For example, $P(y1 = 1|x1 = 1, x2 = 2) = 1$, and $P(y1 = 0|x1 = 2, x2 = 2) = 1$, and so on. This approach represents $x1$ and $x2$ as separate dimensions, and $y1$ is computed as a function of $x1$ and $x2$ jointly. There is no recoding, the elements of the task retain their identity, and the structure is preserved.

It might be asked why this matters, given that the output function can be computed by recoding. The answer is that computing output functions is not the only task of cognitive processes. Sometimes the structure of the task needs to be represented, so as to provide a mental model of the concept embodied in the task. A person who performed the task in C&T's Table 1 by recoding as in Table 2 would not have an explicit, accessible, and transferable representation of the concept, and might not recognize another task as an analog of the task in Table 1. An example of this occurs with conditional discrimination, where learning a unique configuration for each cue-background combination (directly analogous to the recoding in Table 2) provides no basis for transfer to an isomorphic conditional discrimination, because the elements lose their identity, and mappings from cue and background to response are not preserved. Thus there is a potential tradeoff between recoding and accessibility of structure.

The procedure of keeping the input dimensions distinct, and processing them jointly, as indicated above, preserves the structure of the original input. It is subject to capacity limitations, however. Elsewhere we have argued (Halford 1993; Halford et al., submitted) that complexity is best measured by dimensionality, which corresponds to the number of interacting variables in a task. The task in C&T's Table 1 is three-dimensional (it comprises three variables), whereas the task in C&T's Table 3 is four-dimensional. Processing load increases with dimensionality, and adult human processing capacity appears to be limited to approximately four dimensions processed in parallel. Therefore tasks such as those in Tables 1 and 3 could be processed without recoding, but tasks with more dimensions probably could not, because of processing capacity limitations. Thus there is a potential tradeoff between processing capacity and the need for recoding.

Informed learning and conceptual structure: Putting the "birdness" back in the bird

Kenneth Kurtz

Department of Psychology, Stanford University, Stanford, CA 94305. **kurtz@psych.stanford.edu; www-psych.stanford.edu/~kurtz**

Abstract: The computational notion of "trading spaces" is highly relevant to the psychological domain of categorization. The "theory" view of concepts can be interpreted as a recoding view. A design principle for exploiting learned recodings in order to handle the type-2 problem of forming sophisticated concepts is outlined.

Clark & Thornton (C&T) develop an important theoretical claim about learning. The need to recode input data in a manner that reveals higher-order, task-relevant regularities is a computational constraint that ought to be brought to bear in the development and evaluation of psychological theory. Concept formation is one domain that conforms to C&T's notion of type-2 learning. As it happens, the limits of statistical learning is a topic under considerable debate in the psychological literature.

The "theory" view of categorization addresses the issue of the limits of uninformed learning. Proponents of the theory view contend that feature frequencies and correlations underdetermine actual concepts (Murphy & Medin 1985). The theory view portrays accounts built on similarity-based comparison and feature-based representation as either insufficiently constrained or insufficiently powerful to account for categorization (see Goldstone 1994). Murphy and Medin (1985) argue that intuitive theories are needed (1) to explain the coherence of concepts, (2) to provide constraints on what the relevant features are, and to show (3) how these features should be weighted for importance, and (4) what tolerances should be allowed in their values.

According to theory-based categorization, an input activates underlying principles (Murphy & Medin 1985) which serve to link data to world knowledge. Categorization thus consists of knowledge-guided processes such as reasoning with causal contingencies, applying rules, and instantiating schema-like knowledge structures in order to generate an account, rather than a best match, for the input. While the critique of similarity as an explanatory construct has proven powerful, the theory-view has been undermined by the lack of a psychologically plausible models of the nature and role of intuitive theories.

Wisniewski and Medin (1994) have suggested that world knowledge mediates the instantiation of abstract, intermediate properties of the input which in turn determine category membership. This notion of knowledge as something other than data remains ill-

defined and problematic. However, such knowledge is presumably somewhere to be found in the space of functions (too large to search effectively) of possible recodings for solving complex type-2 mappings (C&T). If hard learning can be informed by prior recodings, then perhaps complex categorization can be accomplished without quitting the tangible realm of experience and data. Avoiding this dualism of epistemological content is effectively the same goal as avoiding "heavy duty nativism" (sect. 4, para. 1). Appropriate computation and management of the data set can be sufficient to reveal the "theory" hidden in the data, that is, to discover what is key for accomplishing a task. Thus, "theories" are the basis for informed recoding of the data in a manner that captures critical properties which are not visible or discoverable by unguided search. But how can such a theory be gleaned from computation over available data?

Beyond perceptual similarity, groups of exemplars cohere as conceptual classes on the basis of commonalities in their relevance and usefulness to people. Exemplars converge and diverge in the roles they play relative to particular tasks and goals. This variance provides a critical set of constraints on the organization of concepts. Kurtz (1996) presents evidence for differences in rated overall similarity of pairs of novel items depending on prior classification of the items.

Task-based commonalities can be used as the targets for a backpropagation network (Rumelhart et al. 1986) that constructs internal representations to perform complex mappings from perceptual cues to conceptual targets. These systems have the property of transforming representational space to construct a functional similarity whereby like internal representations lead to like outputs (Rumelhart et al. 1995). Such a mapping from perceptual inputs to abstract classes can be integrated with an auto-encoding system that discovers structure in order to learn to reconstruct the input information at the output level. The two mappings can be superimposed so that the recoding occurs over one shared set of hidden units. Through such learning, objects that do not appear alike yet possess useful conceptual commonalities will be represented more closely.

C&T suggest that language itself is a form of preserved recodings. Language serves as a framework for the organization of entities into classes that are useful for understanding and communicating about the environment. Such lexical classes can provide a set of targets that complement and constrain the structure derived from perceptual regularities in the environment. Another mode of task-based or functional re-representation arises from the domain of problem solving or goal attainment. As particular needs or wants arise across experience and are resolved by particular objects in the environment, a backpropagation network can be trained to learn this mapping from perceptual data about objects to the roles the objects play or the goals they fulfill. This complementary mapping can also be integrated with the auto-associative system. Accordingly, properties relevant to the interaction between organism and environment are integrated into the conceptual representation. Relevant knowledge is merged into the same content as the perceptually grounded data.

Task-driven, linguistic, and perceptual bases of similarity can be assimilated in an underlying conceptual representation of shared hidden units. The features that comprise this representation are recodings of the input mediated by the multiple modes of appropriate generalization which interactively constrain and enrich the emerging representations.

Armstrong et al. (1983) argue that the concept of bird is greater than the sum of bird attributes; that is, the critical quality of "birdness" is notably missing from such a representation. The present proposal is to put the "birdness" back in the bird through the use of recodings built up interactively across multiple mappings. Such rich concepts are organized by distance in a functional similarity space that is relevant to actual concept use and is constructed according to the design principle of parallel mappings through a shared recoding space.

The dynamics of cumulative knowledge

David Leiser

Department of Behavioral Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel. **dleiser@bgumail.bgu.ac.il**
www.bgu.ac.il/beh/leiser.html

Abstract: Exploiting existing representations implies tapping an enormous domain, coextensive with human understanding and knowledge, and endowed with its own dynamics of piecewise and cumulative learning. The worth of Clark & Thornton's proposal depends on the relative importance of this dynamics and of the bottom-up mechanism they come to complement. Radical restructuring of theories and patterns of retrieval from long-term memory are discussed in the context of such an evaluation.

Clark & Thornton's (C&T's) basic point is akin to the observation that finding optimal coefficients for multivariate linear regression is straightforward, whereas devising an appropriate nonlinear model is much harder. To do so effectively, existing knowledge must be exploited. While it is difficult to take issue with this, the full significance of the observation must be realized: no mere tactical recommendation, it implies tapping an enormous domain, coextensive with human understanding and knowledge; a domain, moreover, endowed with its own epigenetic dynamics of piecewise and cumulative learning. If it turns out that most of the novelty required to solve a type-2 problem is ultimately based on that dynamics, then it is misleading to present the enjoyment of the fruits of quite different learning mechanisms as a trick to complement uninformed learning.

Exploiting existing knowledge. C&T contend that a purely bottom-up approach is not enough, and that solving type-2 problems requires the exploitation of previous learning. This could mean transferring knowledge across connectionist systems, a notion of great potential importance, but still in its infancy. Instead, they present a range of dissimilar approaches, all considered equivalent from their very abstract viewpoint: to maximize the role of achieved representation.

But how does that knowledge develop? It is hardly itself a product of type-1 bottom-up learning. Constructions at the conceptual and theoretical level obey coherence and coordination requirements (Thagard 1989) and assume a much richer representational language. The idea was first propounded by Piaget (1974) under the name of reflective abstraction and cognitive phenocopy: what is acquired at one level must be reconstructed and reorganized at another. Piaget's views are being rediscovered in various contexts. According to Keil (1994), there is no obvious mapping from the kinds of environmental regularities most salient to simple associative systems operating on perceptual primitives to sets of explanatory beliefs. This discrepancy arises for a range of reasons. One is the richer representational language. Another may stem from restructuring. Consider cascade correlation: successively generated hidden variables are akin to epicycles; they only explain residual variance (Schultz et al.). Theories, however, may change radically. Some forms of learning may give rise to a weak restructuring, involving the accumulation of new facts and the formation of new relations between existing concepts. Others involve a radical restructuring that includes changes in core concepts, structure, and the range of phenomena to be explained (Carey & Spelke 1994; Gold 1987; Vosniadou & Brewer 1987). This phenomenon is not necessarily incompatible with Clark and Thornton's view, since no actual restructuring needs to take place. One possible interpretation is that the new conceptual structure grows alongside the old one, and eventually comes to replace it (Chi 1992); this interpretation could be handled by a modular connectionism. My point here was to indicate the types of phenomena that may occur at the conceptual level because of the latter's mode of development would be very different from developments that might take place without it.

Retrieval propensities. Another point to consider is organization and retrievability in long-term memory. Here, too, there is ample room for cumulative progress that would then dominate

bottom-up factors. As experience accrues, the store of potentially relevant information grows and must do so in a way that – with an appropriate retrieval mechanism – might guide novel problem solving. Research on analogy suggests that retrieval relies mostly on superficial features and hence tends to stay within domain boundaries. Only as subjects become experts do they develop the abstract tools that enable them to represent the essence of new problems and to apply relevant schemes learned in the past. This suggests that bottom-up and local learning is superseded by more conceptually based general understanding, but only for domain experts (Gentner et al. 1993; Goswami 1991; Reeves 1994). It seems accordingly that the pattern of retrieval propensities from long-term memory does not contribute much to bottom-up learning.

In sum, I have tried to illustrate how the proposal by Clark & Thornton should be critically evaluated. Their basic point is almost trivially true. Assessing its significance demands an appraisal of the relative importance of other mechanisms involved. In the event, I have concluded that the complexities of theory formation should be acknowledged, that restructuring is not necessarily a problem, and that the structure of retrieval from long-term memory does not raise any particular difficulty except in experts.

Extracting higher-level relationships in connectionist models

Gary F. Marcus

Department of Psychology, University of Massachusetts, Amherst, MA 01003. **marcus@psych.umass.edu; www-unix.oit.umass.edu/-marcus**

Abstract: Connectionist networks excel at extracting statistical regularities but have trouble extracting higher-order relationships. Clark & Thornton suggest that a solution to this problem might come from Elman (1993), but I argue that the success of Elman's single recurrent network is illusory, and show that it cannot in fact represent abstract relationships that can be generalized to novel instances, undermining Clark & Thornton's key arguments.

Clark & Thornton (C&T) provide a compelling argument that statistical learners face problems in learning higher-level relationships. Less compelling, though, is their proposed solution, "achieved representational spaces." The centerpiece of their argument is the incremental version of Elman's (1993) model, which they and Elman argue is learning something about grammatical dependencies like subject-verb agreement.

But has Elman's model truly abstracted a higher-order regularity that is not attested in the input? Elman himself does not show this, since the only quantitative comparison he provides is "the degree to which the network's predictions match the . . . probability distributions of the training data."

In fact, Elman's model depends entirely on the statistics of lexical concurrence, never deriving abstract higher-order relations. I discovered this by conducting network simulations that contrasted statistical and relational information (Marcus 1996a). For example, in one experiment, I trained a version of Elman's model on sentences constructed from the grammar *an X is an X,* using a set of twenty different instances of *X* (rose, duck, iguana, butterfly, etc.). What happens if we test how this generalization is applied to a novel word? Given the sentence fragment *a dax is a* _____, humans readily predict the continuation *dax.*

Elman's model behaves quite differently: while it easily learns all of the training sentences, it is unable to extend the abstract underlying relationship to the novel word *dax.* This failure is robust, unaffected by the number of hidden units, the number of hidden layers, the number of training examples, or the sequence in which those training examples is presented.

The reason the network fails to extend the abstract relations lies in its roots as a statistical approximator: within the training corpus the conditional probability of the word *dax* appearing in the frame *a . . . is a* _____ is zero. The model merely mimics those conditional probabilities; it is unable to step beyond those statistics and derive the higher-level abstract relation, substantially weakening C&T's central arguments.

Another problem that might actually demand that the learners go beyond mere input statistics has been the focus of detailed empirical inquiry. My colleagues and I have argued that generalizations of default linguistic inflection (*cat-cats, walk-walked*) are not closely correlated with statistical properties (e.g., Marcus et al. 1995). For example, the German plural *-s* is generalized in much the same cases as the English *-s,* even though the German plural *-s* is required by less than 10% of German nouns.

Most of the dozen or so connectionist models of linguistic inflection – because they are strongly driven by input statistics – face difficulty in explaining these data. Interestingly, one connectionist model accounts for some of these data (Hare et al. 1995), but only by resorting to several innate architectural details of the sort that C&T scorn as "profligate," including an innate, localist node dedicated to the regular *-ed* past tense morpheme, innate inhibitory links between the *-ed* output node and all the irregularly inflected vowel outputs, and innate excitatory links between *-ed* and the unchanged vowels that typically occur in regular past tense forms. Unfortunately, without innate rewiring, the model is unable to learn blends like *sleep-slept* or inflectional systems in which the default is any morpheme other than the model's innately wired *-ed* (Marcus 1996b).

In sum, C&T are right to raise the issue of how connectionist models can generalize beyond simple statistics, but the solutions they offer are too limited. Will networks that lack innate resources ever solve these sorts of problems? Although C&T scorn nativism they give no arguments against it; meanwhile their argument that nonnativist connectionist models could solve these problems doesn't go through.

Data coding takes place within a context

Daniel Memmi

LEIBNIZ-IMAG-CNRS, 46 avenue Felix Viallet, 38000 Grenoble, France. **memmi@imag.fr**

Abstract: Recoding the data for relational learning is both easier and more difficult than it might appear. Human beings routinely find the appropriate representation for a given problem because coding always takes place within the framework of a domain, theory, or background knowledge. How this can be achieved is still highly speculative, but should probably be investigated with hybrid models.

Clark & Thornton's (C&T's) target article is important and timely, expressing clearly and precisely what practicing connectionists have experienced and suspected for some time. The difficulty of learning relational regularities from raw data has indeed been a tacit limit of connectionist modeling, and making this explicit is truly useful work. This can also be viewed as a more general question: purely statistical learnings' difficulty in drawing away from the immediate data and building hierarchies of features and representations that could be reused beyond the narrow context of a single task.

We suspect, however, that the problem is both easier in practice and even more complicated in theory than C&T make it out to be. As a practical matter, the difficulty is not so serious as it might first appear. Engineers and scientists routinely find the appropriate features and the right coding to make a problem tractable. From everyday cognition to scientific research, this is a modeler's daily bread, and it is mostly a question of domain knowledge and experience. As C&T rightly observe, the real issue concerns why recoding is in fact so common and so successful.

The theoretical answer, however, must be much more complex than the panoply of ploys C&T discuss. For not only is the space of

possible recodings open-ended, but there might be different solutions to the same problem. For example, even a simple parity problem admits of several solutions: data could be accounted for by more complex functions than mere parity. On a higher plane, several scientific theories often compete to explain phenomena in the same domain: wave versus particle explanations of light, historical versus structural accounts of language, and so on. Each interpretation selects very different basic features and properties.

In fact solutions are found and chosen within the framework of a whole domain, system or theory, and not in a vacuum. So recoding always takes place in a given context, which can be crucial in guiding and constraining the search for adequate representations. Overall economy, generality, usefulness, coherence, and clarity may then be taken into account. (Thus, in the absence of further data to the contrary, simple parity is a better explanation than a more arcane function.) In this way we try to build (fairly) coherent interpretations of our world, in everyday life as well as in science. The relevant context can range from perception and a background knowledge of the practical world to the most formal cultural constructs of science and technology.

Let us again take up the example (from sect. 3) of a small robot trying to learn the real size of objects from their distance and apparent width. This is indeed a relational learning problem, but one which should appear much less puzzling (though not any simpler) in the context of real biological development. Achieving perceptual constancy in the face of highly variable input is a very general problem (including size, color, and shape as well as sounds), and a prerequisite for categorization and object recognition. Object constancy seems to be partly innate, partly acquired, but the point is that it probably does not develop independently for each feature, modality, and task. There is a general tendency at work here for a whole class of cognitive systems, and it remains to explain how it came about.

Similarly, Elman's (1993) innovative and elegant work on connectionist grammar acquisition from a corpus of sentences is probably only a part of the whole story. Language is acquired in real-world settings, where the semantic and pragmatic function of words is quite obvious (if not always clear). Language development also goes hand in hand with categorization and concept formation. Thus, a basic linguistic distinction such as noun/verb might be worked out as much from a semantic object/event dichotomy as from statistical regularities in the language input. Again, the overall context of the task could well constrain the search for possible solutions.

C&T seem to allude to such a contextual view when they write about the importance of language and culture, but they do not elaborate about actual mechanisms. It is true that very little is known about how knowledge structures could emerge and be progressively reorganized within whole domains and cognitive systems. Apart from Piaget's (1971) grand but frustratingly vague developmental framework, there is little theory available, whether in cognitive psychology or in Artificial Intelligence (Leiser 1995). Work in symbolic machine learning does not help much, because it avoids the fundamental problem of how to choose basic features and properties. And Karmiloff-Smith's (1992) brave attempt in psychology remains imprecise.

Yet we feel that this avenue of research cannot be pursued for the time being with purely connectionist techniques because these are still not powerful enough. The complexity of the knowledge structures and mechanisms necessary to account for in-context learning would most probably require hybrid systems. Only with such symbolic-connectionist models could we hope to deal with the systematic interaction between prior knowledge and the acquisition of new regularities (Wilson & Hendler 1993; Sun & Bookman 1996).

Of ants and academics: The computational power of external representation

Jon Oberlander

Human Communication Research Centre, University of Edinburgh, Edinburgh EH8 9LW Scotland. **J.Oberlander@ed.ac.uk**
www.cogsci.ed.ac.uk/people/jon/

Abstract: Clark & Thornton speculate that intervening in the real world might be a way of transforming type-2 problems into type-1, but they state that they are not aware of any definite cases. It is argued that the active construction of external representations often performs exactly this function, and that recoding via the real world is therefore common, if not ubiquitous.

Towards the end of their discussion, Clark & Thornton (C&T) observe that in dividing up the cottage cheese,

> we actively manipulate the real world so as to translate the abstract mathematical problem into a form that exploits the specific computational powers of our visual systems. We do not know of any concrete cases in which such physical interventions act so as to transform a type-2 search into some more tractable form, although it may prove fruitful to examine cases in which color-coding, chemical trails, and so on, are used to simplify recognition and tracking. (sect. 4, final paragraph)

It is true that an ant's chemical trail is a physical intervention; but so too are the many different ways in which people mark surfaces within their environment. Thus, writing text, drawing diagrams, sketching maps, filling in forms, drafting a histogram – even typing on a computer keyboard – are all ways of constructing external representations and adding new structure to our environment, thereby guiding our future behaviour.

In particular, such external representations – of which diagrams are a paradigm – can play a powerful role in assisting human inference, for a number of reasons. First, it is commonly observed that they reduce memory load, and this is especially true if the diagram is incrementally modified during problem-solving, thereby storing the results of reasoning episodes (cf. Larkin & Simon 1987). Second, as C&T note, our visual system has special-purpose hardware that solves certain problems extremely rapidly (cf. Funt 1980). Finally, when we make marks on a surface, the physical nature of that surface introduces a whole set of additional constraints. For example, following Bishop Berkeley, we can see that a triangle drawn on a sheet of paper has to have a specific set of angles and line lengths – it cannot, without further conventional interpretation, be a "generic" triangle. Similarly, if I draw a set of entities in a blocks world diagram, there will be a specific number of blocks. Stenning & Oberlander (1995) investigated such cases, and concluded that external representations trade in these constraints – which prevent complex propositions from being expressed – for computational efficiency. The point is just this: a surface in the world has certain affordances, which force representations constructed upon it to precompile a huge number of inferences, thereby saving effort later (cf. Lindsay 1988).

This precompilation implies that building an external representation effectively involves recoding the problem space. In particular, there are plenty of instances of representation construction that actually convert type-2 problems into type-1.

Funt's work is one example, but it seems plausible to make the general suggestion that everyday data visualisation tools – such as tables, graphs, and charts – perform exactly the type-2-to-type-1 recoding function; their fitness for this task explains why good visualisation tools survive, alongside natural language. To see that this is so, consider the training data in Table 1, which C&T discuss in (sect. 2, paras. 12–14). With two input variables ($x1$ and $x2$), and one output ($y1$), the problem as presented in Table 1 is characterised as type-2. C&T suggest that if we recode the input "with a single variable whose value is just the difference between the original variables" then we will quickly derive useful probability statistics within the recoded data.

Different visualisations offer different recodings, and their utility can be evaluated by eye. Take C&T's Table 1: it can be

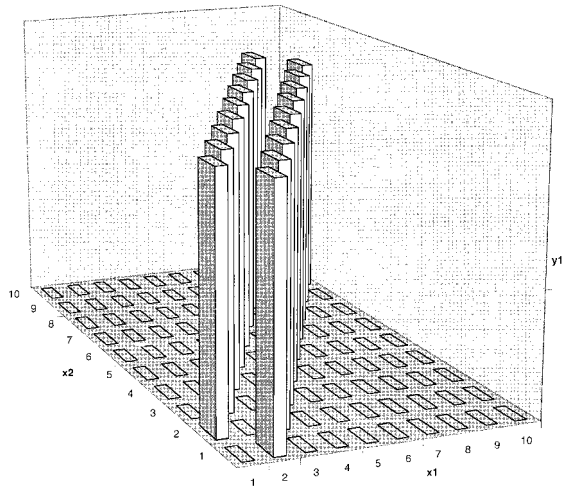

Figure 1 (Oberlander). Recoding Clark & Thornton's Table 1 by visualising a space.

recoded as a two-dimensional table, with $x1$ and $x2$ in rows and columns, and $y1$ as values for cells in the table:

		x2		
		1	2	3
x1	1	0	1	0
	2	1	0	1
	3	0	1	0

Because we have formed a Cartesian space, the diagonal of 0s is immediately salient, and one way of viewing the 1s is as flanking the diagonal; hence we can *see* that $y1 = 1$ where $x1$ and $x2$ differ in value by 1, and $y1 = 0$ otherwise. The point is made even more vivid if we render a 3D bar chart of a larger training set, extended so as to remain compatible with C&T's generalisation (see Fig. 1). Although C&T do not use either of these visually oriented recodings on their Table 1 data, it is clear from the target article that they must be academics who subscribe to the power of visualisation; otherwise, we cannot explain why their paper does contain a line-graph (Fig. 1 of the target article).

Academics – and other people – habitually mark the environment in the process of problem-solving; because the environment has certain affordances, recoding happens as a matter of course; and because our visual system is remarkably good at detecting certain features of the environment, higher-order relationships leap out of the now modified environment.

The moral of this story is that the study of diagrammatic reasoning, data visualisation – and external representation use more generally – should provide a rich source of "concrete cases of physical interventions which serve to transform type-2 search into some more tractable form." Ants certainly are interesting, and studying their chemical trails should provide insight into the way that marking the environment helps to reduce computational load. But in the long run, academics may prove even more interesting than ants.

Old ideas, new mistakes: All learning is relational

Stellan Ohlsson

Department of Psychology, University of Illinois at Chicago, (M/C 285), Chicago, IL 60607-7137. **stellan@uic.edu**

Abstract: Learning is the acquisition of knowledge, not of input/output mappings. The distinction between statistical and relational learning, as Clark & Thornton define those terms, is not useful because *all* human learning is relational. However, prior knowledge does influence later learning and the sequence in which learning tasks are encountered is indeed crucial. Simulations of sequence effects would be interesting.

Clark & Thornton (C&T) provide solid evidence that some connectionist systems cannot learn to recognize the difference between large and small objects. This result is not accidental; the authors see "a robust pattern of failure" on this and other, equally simple tasks (sect. 2). Somewhat surprisingly, this failure is taken as a basis for far-reaching speculations about the nature of mind, the function of culture, and the course of hominid evolution.

The conceptual bridge between the failure and the speculations is a distinction between "statistical" and "relational" learning tasks, plus the three ideas that (a) the learning mechanisms in our heads can only handle statistical learning, (b) relational learning tasks can only be solved if prior learning reduces them to statistical tasks by enabling the learner to recode the relevant information, and (c) the effect of prior knowledge on later learning implies that the sequence in which learning tasks are encountered is crucial.

This is an interesting argument. Unfortunately, C&T define "the process of learning . . . as the attempt to acquire a target input/output mapping" (sect. 2). This is, of course, exactly what learning is not. The idea that what is learned is an input/output mapping (or a set of stimulus-response connections) was abandoned in the 1950s because people began taking the generativity of human cognition seriously (Chomsky 1959).

It appears that the case against behaviorism bears repeating: People adopt perspectives and form beliefs; they remember the past, visualize the future, and imagine the impossible; they have intentions and goals; they plan and make decisions. People act differently in one and the same situation, depending on what they believe and want. In short, human behavior is generated centrally, not peripherally; action is adaptive, not reactive. Hence, learning cannot be understood as the acquisition of an input/output mapping, regardless of type and complexity. There is no reason to revive the behaviorist view.

We are thus faced with an interesting argument formulated within an uninteresting framework. Can we extract the former from the latter? Are there any real world examples of C&T's concept?

There are abundant examples of the idea that prior learning is a major determinant of new learning, because this is one of the oldest ideas in cognitive research (Ausubel 1963). Effects of prior knowledge have been documented in text comprehension, perception, problem solving, and learning (Eysenck & Keane 1995). The idea that prior learning enables the learner to recode information is also well documented, although it is usually referred to as "chunking" rather than "recoding" (Miller 1956). Examples of the importance of sequential order of learning tasks are particularly easy to find in educational contexts: vocabulary comes before reading, algebra before calculus, and so on (Gagne 1962).

So far, so good. However, when we get to C&T's own proposal – namely, that there are two qualitatively different types of learning and that an instance of the computationally more demanding type can be reduced to a sequence of less demanding ones – then I'm stuck. I am unable to think of a single convincing example. There are many distinctions between qualitative types of learning (empirical vs. theoretical, declarative vs. procedural, explicit vs. implicit), and prior learning of one type can certainly facilitate learning of another type. Prior knowledge helps, but it does not help *in the particular way* proposed by Clark & Thornton. For

example, a good vocabulary does not "recode" a text so as to reduce the computational complexity of reading, nor does knowledge of algebra "recode" calculus problems in such a way as to reduce the computational complexity of learning calculus. In general, prior knowledge does not help by changing a learning task from one type (with high computational complexity) to another, qualitatively different type (with lower computational complexity).

I believe that C&T have it exactly backwards: They claim that statistical learning "is really all we currently know how to achieve" (sect. 5). On the contrary, it seems to me that all human learning is relational learning in the sense that prior knowledge is used to interpret ("recode") the learning situation and new learning proceeds on the basis of the interpretation. Statistical learning is all that *connectionist learning systems* know how to do; this has no implications for human learning.

In short, lifted out of its behaviorist framework and shorn of its pseudo-mathematical trappings, the distinction between statistical and relational learning has nothing going for it; but the emphasis on the role of prior knowledge and the sequential order of learning tasks is right on target. Although there are simulations of the effects of prior knowledge on learning (e.g., Ohlsson 1996; Ohlsson & Rees 1991), sequence effects have not been studied in depth. Simulating such effects in a real learning scenario would require a lot of work. Perhaps C&T will get around to it once they finish teaching their current system to recognize the difference between large and small objects.

Neural computation, architecture, and evolution

Paul Skokowski

McDonnell-Pew Centre for Cognitive Neuroscience, Oxford University, Oxford, OX1 3UD, England. **paul.skokowski@psy.ox.ac.uk**

Abstract: Biological neural computation relies a great deal on architecture, which constrains the types of content that can be processed by distinct modules in the brain. Though artificial neural networks are useful tools and give insight, they cannot be relied upon yet to give definitive answers to problems in cognition. Knowledge re-use may be driven more by architectural inheritance than by epistemological drives.

The nativism/empiricism debate is intriguing, but perhaps a little overblown for high-level cognition. It is clear that much of the architecture of the brain is pre-determined. The various regions studied by neuroscientists and psychologists and the neural pathways between and through them do not vary substantially from subject to subject. This much speaks for nativism. Yet people (and animals) manage to learn new things, including some very difficult cognitive tasks indeed, throughout their lifetimes. Despite the evidence that its overall architecture appears to be genetically determined, the brain still turns out to be plastic. This is because of the modifiability of synaptic connections within this predetermined structure. A main contender for how this feat is accomplished at the neural level is Long Term Potentiation, or LTP, yet if we cut certain neural pathways in the right places, in man or beast, some types of cognitive function may never be recovered. People with certain kinds of localized brain damage become blind. No other region of the brain can compensate and allow them to see, despite the massive number of interconnections remaining in the brain. Plasticity, therefore, has its limits.

The point is that the brain is partially predetermined, and partially plastic. It has inherited certain tricks that help it to solve certain problems. Architecture and direct wiring is one very powerful trick our evolution has provided us. Few would deny that the visual cortex has evolved to solve visual tasks. But having said that, learning is still required for proper vision. Children learn to focus on objects and distinguish actual sizes of objects that occupy equivalent solid angles of their visual field. Deprivation of visual stimuli at certain key developmental stages in cats leads to blindness. The brain is a bag of tricks, to use Dennett's (1992) terminology, to be sure, but it is a bag of tricks that still requires learning – plasticity – at numerous levels in order to use these tricks for the benefit of the organism.

Clark & Thornton (C&T) now tell us that there are other tricks in store. We are told that cognitive systems need an informed search to solve certain problems. There is some merit to this claim, but I wonder about their approach. We are told that neural networks, or animats with neural network drivers, both having restricted classes of neural net architecture, cannot solve certain types of problems. We are then reminded of Elman's (1993) neural network studies in which early training on one type of input corpus, followed by training on another type of input corpus, yields encodings adequate to solve a complex problem in language. These examples are meant to support the claim that, in order to achieve certain tasks, we (successful cognitive systems) must go beyond finding simple correlations in sensory input, and manipulate higher order representations. Perhaps for connectionists this is a new discovery (which I doubt). But haven't we known this for some time about biological systems which have been endowed with substantial neural machinery: systems like us?

A hidden assumption in C&T's target article that must be addressed is that what goes for neural networks goes for biological cognitive systems. When a fixed-architecture neural network fails on a certain task, does that really tell us something specific about our cognitive capacities? Or does it tell us something about feed-forward neural networks of that sort of architecture? Perhaps another layer, or simulated annealing, or some other neural network would have ameliorated the situation and allowed a solution of the problem with a neural net approach. What would this imply for C&T's thesis? Though I find such networks to be useful tools and to give us insight into our cognitive life, I remain healthily skeptical about their role as final arbiters on cognition. Having said this, I must agree with C&T that if the brain does avail itself of neural net style computation, then it must be a massively parallel network of such networks.

C&T make the further claim that we have a drive to re-use old knowledge. But look at the brain again. As mentioned above, it is architecturally configured in a certain way. If we, as individuals or as a species, must deal with new environmental contingencies that haven't occurred in our evolutionary past, we must use the tools we have. This much seems trivial. Architecturally speaking, you can't teach an old dog new tricks – with the obvious exception of several thousand years of evolution. Though architecture itself isn't knowledge (content), it constrains the types of content that can be processed by that module. If we use an existing (trained) module in the brain to try to do a new task, then the only tools it will have to work with are its architecture and its current content. I'm not sure if that constitutes a drive: bare necessity may be a more apt description.

Finally, evolution takes time. Language changes much more quickly (witness valley or cyber-speak) than species ever could. C&T are therefore on the right track, in my opinion, in claiming that language adapts to our capabilities rather than the other way around. Language, however, is a peculiar beast, being made up of tokens, meanings, speakers, grammar, and so on. It would take a lot of hard work to complete this story to everyone's satisfaction.

Why computation need not be traded only for internal representation

Robert S. Stufflebeam

Philosophy-Neuroscience-Psychology Program, Washington University, Campus Box 1073, St. Louis, MO 63130-4899. **rob@twinearth.wustl.edu**

Abstract: Although Clark & Thornton's "trading spaces" hypothesis is supposed to require trading internal representation for computation, it is not used consistently in that fashion. Not only do some of the offered

computation-saving strategies turn out to be nonrepresentational, others (e.g., cultural artifacts) are external representations. Hence, C&T's hypothesis is consistent with antirepresentationalism.

Using achieved representations reduces computational labor while also enabling a computational device to solve what would otherwise be an intractable mapping problem. Such is the crux of Clark & Thornton's C&T's representation-dependent "trading spaces" hypothesis. There are two facets of their gloss on the hypothesis about which I shall comment, one pertaining to the representational status of the sorts of "ploys, stratagems, and tricks" (sect. 4, para. 6) they identify, the other pertaining to whether the hypothesis depends on internal representations.

This commentary is motivated by my research into the nature of representation and its relation to computation. Since C&T's hypothesis is predicated on the interrelationship between these notions, let me begin by exploring a bit of the dialectic between computationalists and their opponents.

Although C&T fail to state precisely what they take to be an internal representation – apart from "a recoding scheme" (sect. 3, para. 16) – it is clear that they feel internal representations play a crucial role in biological computational processing. They are not alone: *computationalism* and *representationalism* underlie virtually every naturalistic attempt to explain how the mind/brain works. The former is the view that a computational framework is essential to explain the workings of "intelligent" systems. The latter is the view that computational systems require a sophisticated system (or "medium") of internal representations, without which they cannot compute (Fodor 1975, p. 27). So, regardless of whether one's preferred computational framework is symbolic or nonsymbolic, the status of internal representations in explanations of computational processing seems to be secure. Or so we have been conditioned to believe.

The received view of computationalism is not without its gainsayers. Although attacks against it come in a variety of guises, most (though not all) varieties of anticomputationalism are explicitly antirepresentational. Some of the more significant critiques include: (1) attacks against a computational framework being a plausible framework within which to explain cognition (Port & Van Gelder 1995; Van Gelder 1995); (2) arguments against the notion that (biological) computation presupposes a medium of internal representations (Stufflebeam 1997, Chs. 3–4); (3) attacks against the biological plausibility of cognitivism and symbolic computation (Dreyfus 1992; Searle 1990); (4) attacks against the efficacy of computational simulations as a basis for explanations of how the mind/brain works (Fodor 1995; Searle 1990); (5) arguments in defense of situated action (Agre 1993; 1995; Brooks 1991); and (6) antirepresentational arguments regarding the role and status of "distributed representations" in explanations of PDP (Stufflebeam 1995). If anyone is familiar with these attacks, it's Andy Clark (Clark 1996; Clark & Toribio 1994).

Does any computation-saving ploy, trick, or stratagem qualify as a representation? Some do – for example, "public language" and certain cultural artifacts (sect. 4, para. 9). Because to use public language is to use achieved representations to reduce the computational complexity regarding (at least) the transmission of knowledge, since public language is clearly representational, it exemplifies C&T's hypothesis. But "real-world actions" are supposed to do that as well (sect. 4, para. 15; sect. 5., para. 4). Here's the rub: C&T are careful *not* to call such computation-saving ploys "representations," though they *do* feel real-world actions are consistent with their hypothesis. As such, it is odd that C&T are insensitive to the antirepresentationalist arguments coming from proponents of situated action.

More important, C&T's hypothesis does *not* seem to depend on trading internal representation for computation, as is their claim. Instead, it seems to depend rather on trading *something* for computation, even if that something is an *external* representation (as is the case with cultural artifacts and public language utterances). It may still be the case that "the computationally weak will inherit the earth" (sect. 5, para. 5). But, one could argue, it is the *external* representations that make the computationally weak "representationally rich enough to afford it" (sect. 5, para. 5). And since real-world actions *also* get traded for computation, their hypothesis is far less representation-dependent than Clark & Thornton seem to realize.

Prospects for automatic recoding of inputs in connectionist learning

Nicolas Szilas and Thomas R. Shultz

Department of Psychology, McGill University, Montreal, Quebec, Canada H3A 1B1. **nicolas@lima.psych.mcgill.ca; shultz@psych.mcgill.ca; www.psych.mcgill.ca/labs/lnsc/html/lab.-home.html; www-leibniz.imag.fr/reseaux/szilas/szilas.html**

Abstract: Clark & Thornton present the well-established principle that recoding inputs can make learning easier. A useful goal would be to make such recoding automatic. We discuss some ways in which incrementality and transfer in connectionist networks could attain this goal.

Clark & Thornton's (C&T) key point, that recoding inputs can make learning easier, is well established and widely known. The merit of their approach is to tackle the problem instead of avoiding it, by showing negative results obtained without recording and providing a general overview of how manual recoding might work.

The next challenge would be to make such recoding automatic. Unfortunately, a mechanism for automatic recoding is not proposed by C&T. Ideas that could serve as the basis for automatic encoding can be found in studies of techniques for making learning easier, only some of which are mentioned by C&T.

C&T wisely cite Elman's (1993) work on incremental learning. There are a few studies showing that learning is sensitive to the sequencing of tasks (Cloete & Ludik 1993; Szilas & Ronco 1995; Tetewsky et al. 1995). However, if the sequence of tasks is not very carefully chosen, learning can be impaired.

A growing subfield in connectionism concerns the study of knowledge transfer. Some of these studies show that a common hidden layer can be shared by several tasks, in either simultaneous (Caruana 1993) or sequential (Baxter 1995) learning.

Transfer from one task to another can be useful only if the tasks are related in some important way. Otherwise, the two tasks may merely interfere with each other. If a representation is to be retained and reused without interference, it should perhaps be frozen. This is what is achieved by an algorithm like cascade-correlation (Fahlman & Lebiere 1990), which builds a hierarchical layered structure in which input weights to hidden units are no longer adjusted once the unit has been installed.

Simulations confirm that cascade-correlation networks are less susceptible to retroactive interference and make better models of human learning on sequential learning tasks than more conventional back-propagation networks that do not freeze hidden units (Tetewsky et al. 1993). In these simulations, transfer is achieved by learning connections from an old structure to a new one, whereas C&T seem to discard this possibility unless the earlier subnet is copied. With freezing of input-side weights, subnets are not used up, but simply used.

The constraint that "the dimensionality of the inputs is identical for both the original task and any later ones" can likewise be overcome: once again, use connections! A three dimensional input can be connected to a four dimensional input by a set of weighted links. Furthermore, in referring to Karmiloff-Smith's (1992) Representation Redescription, C&T seem to identify recoding with abstract redescription. Even if abstract redescription does exist, the foregoing examples show that the reuse of knowledge can occur without abstraction.

C&T stress the importance of using old representations to facilitate new learning, in effect, trading representation for computation. However, it is worth noting that in open-ended sequen-

tial learning there may be as many representations as there are learning episodes. Consequently, using achieved representations implies searching among them, and such search does require computation. Psychological studies of analogical reasoning show that even when people have relevant knowledge, they may not be able to use it without extensive hints (Gick & Holyoak 1980; 1983). Because such search is not trivial and is often unsuccessful, C&T's space trading maneuver is not without potential problems. The more representations that are available, the more extensive the search computation is likely to be.

Relational problems are not fully solved by a temporal sequence of statistical learning episodes

A. Vinter & P. Perruchet

L.E.A.D., C.N.R.S., University of Bourgogne, 21000 Dijon, France.
vinter@satie.u-bourgogne.fr

Abstract: Clark & Thornton's conception finds an echo in implicit learning research, which shows that subjects may perform adaptively in complex structured situations through the use of simple statistical learning mechanisms. However, the authors fail to draw a distinction between, on the one hand, subjects' representations which emerge from type-1 learning mechanisms, and, on the other, their knowledge of the genuine abstract "recoding function" which defines a type-2 problem.

1. Power of statistical learning mechanisms. Much of the interest of Clark & Thornton's (C&T's) target article lies in the fact that it offers a straightforward demonstration of the power of statistical learning mechanisms for solving problems which seem, prima facie, to be beyond the scope of such mechanisms. Empirical support for this conclusion can be found in the recent literature on implicit learning (Dienes & Berry, in press). In an often-cited study (Lewicki et al. 1988) for example, participants were asked to track as fast as possible a long and continuous series of targets appearing apparently at random locations. Unknown to participants, the series was composed of a systematic alternation of two unpredictable and three predictable trials. The discovery of this structure implies that subjects recode the continuous succession of trials into adjacent blocks of five successive trials. The underlying structure of the series remained completely opaque to participants, even after practice, yet performances were better for the predictable trials than for the unpredictable ones. Perruchet et al. (1990) demonstrated that the surprising adaptive performance of subjects in this situation was a direct consequence of a sensitivity to the frequency of occurrence of certain small chunks of two or three trials generated by the rules structuring the series. One could say that subjects solved a type-2 problem after its reduction to a set of type-1 problems.

The analogy between C&T's position and some aspects of the literature on implicit learning may be taken a step further. Perruchet and Gallego (in press) have proposed a theoretical account of implicit learning which shares striking similarities with C&T's claims about the nature and the function of type-1 learning. In this account, implicit learning is devoted to the formation of the "subjective units" shaping the perception of events and objects. Statistical learning mechanisms result in the chunking of information into discrete units, the nature and size of which are a function of the salience of surface features, as well as of the subject's background knowledge and general processing constraints and abilities (active memory and attention mainly). These subjective units emerge from the association of the primitive features that are processed conjointly in an attentional focus, and determine how the environment is attentionally perceived and processed after experience. With training, these units become increasingly independent of the sensory input and hence form internal representations. In line with C&T's position, this account construes the notion of representation as the endproduct of statistical learning mechanisms, making it possible to deal efficiently with problems involving what are a priori powerful computational abilities.

2. Limits of statistical learning mechanisms. Placing C&T's conception of learning within the context of implicit learning research reveals a major limitation of this conception, however. First note that C&T do not distinguish between the formation of achieved internal representations of the world, which permits behavioral adaptation to a given situation, and subjects' knowledge about the structural features of this situation. Let us illustrate this distinction. Each of us can state the direction of the source from which a sound comes. This ability stems from the detection and analysis of subtle differences in intensity or phase between the auditory streams processed by each ear. Consequently, location detection belongs to the class of relational, type-2 problems. The distinction we refer to is between the formation of achieved representations of sound space and the knowledge of the principle which permits these representations, namely, that detection is possible thanks to the relation between the information provided to each ear (Vinter & Perruchet 1994). Now, as should be clear from this example, it makes no sense to endow laymen with knowledge of this principle. The idea of knowledge makes sense here only from the observer's point of view not from the subject's.

In location detection, the coding of relational information is the direct product of hard-wired mechanisms. Our proposal is that the very same logic holds for the recoding provided by type-1 mechanisms of learning. The sensitivity to frequency statistics, and the representation resulting from this sensitivity, must be carefully distinguished from the subject's knowledge of the relational properties embedded in the task. Let us return to the Lewicki et al. situation. We noted that the better performance of subjects on the predictable trials, which apparently indicated that subjects were sensitive to the underlying structure of the series, relied on the sensitivity to the frequency of certain chunks forming the series. The crucial point is that this sensitivity to the surface frequency features gave the subjects no access at all to the underlying structure, for the very reason that the relevant frequencies, although a byproduct of the rules, do not make it possible to infer the rules. Indeed, the rules were concerned with the trajectory defined by two successive locations, whereas the resulting frequency effects captured by the participants were mostly concerned with perceptually salient units such as back and forth movements involving three successive locations. In this situation, it is clear that there is no justification for inferring relational knowledge from improved performance.

3. The need to introduce higher-level processes. We suggest that the solution provided by statistical learning mechanisms to type-2 problems is only a first step in the full course of human learning. The genuine knowledge of the relation embedded in type-2 problems involves processes that C&T fail to consider. In order to gain knowledge about the mechanisms involved in the detection of sound location for instance, scientists need to proceed by reasoning, hypothesis testing, and logical inference. The fact that they are themselves able to detect, as can everyone else, the location of a sound is of no help. In other words, knowledge of the "recoding function" can only be achieved by using processes fundamentally different from those involved in statistical learning. These high-level processes are needed to infer any abstract relation and to integrate it into a coherent view of the world or even to transfer it to another domain. The formation of abstract knowledge implies the use of processes which rely on the specific power of conscious thought. Overall, C&T's suggestion that there is no other type of learning to be had than type-1 learning, needs revision.

Evolution's gift is the right account of the origin of recoding functions

Andrew Wells

Department of Social Psychology, The London School of Economics and Political Science, London WC2A 2AE, England. **a.j.wells@lse.ac.uk**

Abstract: Clark & Thornton argue that the recoding functions which are used to solve type-2 problems are, at least in part, the ontogenetic products of general-purpose mechanisms. This commentary disputes this and suggests that recoding functions are adaptive specializations.

Clark & Thornton (C&T) have enhanced our understanding of the nature of problem spaces with their distinction between type-1 and type-2 problems. Type-1 problems are solved directly. Type-2 problems are solved indirectly via recoding functions which make manifest otherwise hidden regularities in input domains. Recoding reduces type-2 problems to type-1. This is a valuable insight and the statistical framework within which the distinction is formalized is both apt and informative. C&T suggest that type-2 problems are commonplace in a wide variety of domains but are regularly solved despite their computational intractability, which results from the infinity of potential recoding functions.

A central question, therefore, concerns where the recoding functions used to solve type-2 problems come from, and it is here that C&T's analysis is problematic. One possibility, which is currently at the heart of much promising work in evolutionary psychology (Barkow et al. 1992), is that recoding functions are task-specific, evolved adaptations. C&T are clearly unsympathetic to this idea, which they call "heavy-duty nativism" and they accept only that it "is no doubt sometimes plausible" (Abstract).

The reasons for C&T's hostility to the evolutionary solution are not made clear, but they appear to be uncomfortable with the idea of a large number of adaptively specialized type-2 problem solving devices. One of their goals is, therefore, to show how more general-purpose mechanisms and processes might be used in the solution of type-2 problems. Their preferred account of recoding functions builds them via a two-stage ontogenetic process which retains a central role for associative learning. C&T argue that the first stage consists of associative learning over type-1 problem domains which results in the achievement of specific representational states. In the second stage the achieved states can be used more generally on a trial and error basis as potential recoding functions for arbitrary problems which might thus be reduced from type-2 to type-1.

The trouble with C&T's ontogenetic scheme is that it does not solve the problem and hence there is no reason to prefer it to the phylogenetic account of the origins of recoding functions. Let us suppose that associative learning can modify the connectivity of a module or a subnet, as hypothesized by C&T, to realize a specific function which solves a type-1 problem. Let us further suppose that the module can then be made more widely accessible for use by other input domains, perhaps in one of the ways suggested by Rozin (1976). It is hard to see what advantage this confers. The difficulty with type-2 problems is that the space of potentially applicable recoding functions is infinite. All that the first stage of the ontogenetic process can achieve is to make one of these recoding functions available to the problem solver. But unless problem spaces have related structures, the second, trial and error, stage of the process would be of no value, because the probability is vanishingly small that the acquired recoding function would just happen to reduce an independent type-2 problem in a useful way. C&T appear to have no way to avoid this conclusion because they accept that the principle relating achieved representations to problem spaces is chance. "Each such recoding may just happen to reduce a problem that was previously type-2." (sect. 4, para. 6).

C&T's reluctance to accept a phylogenetic account of the origins of recoding functions is all the more curious in the light of their enthusiasm for trading computation for representation. Given that type-2 problems "permeate biological cognition right down to its roots" (sect. 5, para. 1) it is clearly the case that selective pressure would be exerted in favour of mechanisms which instantiated more powerful representations and thus solved problems faster or more accurately than those that did not. It is a computational truism that special purpose machines are faster and more efficient than general purpose machines and it is also evident that natural selection preserves mechanisms which offer selective advantage with respect to specific problems. Evolution's gift of an appropriate set of type-2 problem-relevant recoding biases is exactly what we ought to expect.

Authors' Response

Relational learning re-examined

Chris Thornton[a] and Andy Clark[b]

[a]Cognitive and Computing Sciences, University of Sussex, Brighton, BN1 9QH, United Kingdom; [b]Philosophy/Neuroscience/Psychology Program, Washington University in St. Louis, St. Louis, MO 63130.
chris.thornton@cogs.Sussex.ac.uk; andy@twinearth.wustl.edu

Abstract: We argue that existing learning algorithms are often poorly equipped to solve problems involving a certain type of important and widespread regularity that we call "type-2 regularity." The solution in these cases is to trade achieved representation against computational search. We investigate several ways in which such a trade-off may be pursued including simple incremental learning, modular connectionism, and the developmental hypothesis of "representational redescription."

The target article explores a familiar topic (the limits of simple statistical learning) in what we hope is a rigorous and challenging way. Its motivation was simply the observation that certain types of problem are both frequently solved (by biological learning devices) and yet appear highly intractable from a statistical point of view. These intractable (so-called "type-2") scenarios are ones in which the learner must identify *relations* among raw input elements rather than associations. The puzzle is: how is it possible for limited biological agents to negotiate such statistically impenetrable problem domains? The answer is (we claim) that short of being provided with antecedent search-space-shrinking knowledge (in which case the problem does not arise) the only hope lies in a "bag of tricks" approach that exploits general strategies for pressing maximal effect from those rare cases in which, by chance, a useful re-coding has been found. Re-coding is essential since it is a process that can take a relational property and turn it into a bona fide higher level element in a new space in which previously complex and elusive properties (such as relations between relations) appear as simple patterns (relations).

This thesis, we concede, can seem by turns trivial (of course higher order relational learning is tough!), wildly speculative (surely there are more direct ways of solving this kind of learning problem?), over-technical (did we really need statistics to make our point?), and under-technical (just how precise is the type-1/type-2 distinction anyway?). It is to the great credit of the commentators that, pretty much without exception, they responded constructively, repeatedly underlining our central theme and offering a wealth of useful suggestions and links to other bodies of work. Their responses bear mainly on six issues and we divide our Response accordingly.

R1. Is there still a Grand Ploy waiting to be discovered?

Our claim was that no general algorithm can exist for the systematic discovery of type-2 regularities in unrestricted domains. The most nature can do is to press maximal utility from whatever re-codings are found by chance or by simple search in less problematic domains, or to adjust the problem space itself so as to better exploit the existing biases of the human learning device (as in the Newport [1990] conjectures about morphology). Some commentators, however, proved unable to suppress a laudable optimism and felt (**Berkeley, Haberlandt**) that some more powerful and general mechanism might yet be available. Thus Berkeley, while appearing to be in agreement with much of our thesis, suggests that backpropagation networks using nonmonotonic units can in fact deal with the type-2 parity-generalization scenario which we refer to in our paper. He cites a number of simulation results which back this up. We have no difficulty with this proposal and would only comment that the use of such "nonmonotonic" units equips the learning method in question with an implicit recoding ability and that this just happens to be appropriate for the problem domain he concentrates on, namely parity generalization. Thus Berkeley effectively demonstrates that a suitably biased type-2 method can solve a type-2 problem. Such a demonstration, however, is in no way suggestive of the re-coders grail: a fully general algorithm that achieves type-2 learning whatever the domain.

Several commentators (**Oberlander, Stufflebeam,** and, to some extent, **Kurtz**) suggested that the nearest thing that nature provides to such a general Grand Ploy may be the use (by a lucky few evolved creatures) of a variety of *external* representational systems such as language, maps, graphs, and other kinds of real world structure. This is a powerful and important suggestion, and one that we merely touched upon in our original treatment (see our comments on Dennett in sect. 3 and on the potential role of real world structures and action in sect. 4). We are, however, in full agreement with the idea that external representations play a very major role in empowering biological learning devices (Clark, 1989, Ch. 7; Clark 1997).

We found **Oberlander's** thoughtful and constructive commentary of special help in this regard. Oberlander develops a number of very compelling examples of ways in which we can simplify inner computations by adding structure to the local environment. This is a theme whose time has clearly come, for it is surfacing again and again in recent and influential work on so-called embodied and embedded cognition (see e.g., Hutchins 1995 and Clark 1997), and it is one that we intend to pursue in detail in our future work.

We cannot resist relating a further example, shown to us by Roger Thompson (personal communication), that seems perfectly to illustrate this theme. It concerns the ability of certain chimpanzees (pan troglodytes) to use experience with external tokens to enable them to perform higher order matching tasks that they would otherwise find impossible. The basic result, described at length in Thompson et al. (in press) is that when trained to associate a relational feature of some inputs (e.g., the feature of sameness) with an arbitrary external token (such as a plastic heart), the chimps can go on to learn to perform a higher order task (matching relations *between* relations) that would otherwise defeat them. Thus they become able to judge of two pairs of objects – such as two identical shoes and two identical cups – that the pair of pairs is an instance of the sameness relation at a higher level, that is, sameness in respect of sameness, each pair being itself an instance of the basic relation of object level sameness. This task of matching relation between relations is, we think, a clear instance of a type-2 learning scenario. But one in which, as predicted by **Oberlander,** the existence of external tokens capable of reifying the relations between basic domain elements renders the problem tractable to on-board biological cognition. We here trade externally provided props and structures against expensive and perhaps even intractable episodes of inner computation.

In addition to the dedicated seekers after a Grand Ploy, some commentators suggested useful additional locally effective props and stratagems that might be added to our bag of tricks. **Szilas & Shultz** note the virtues of cascade correlation networks and suggest that a greater use of between network connections may do much to reduce the need for whole network copying and to overcome mismatches of input size during episodes of analogical reasoning. We agree that these and other technical tricks may help explain in detail how codings developed in one domain get to be transferred to others in which they may, at times, reduce type-2 complexity to type-1 tractability. The basic strategy however is still simply the re-use of achieved representation – it is trading spaces just as we envisaged it.

R2. The role of evolution

One contentious move in our original treatment was to avoid reliance on what we (perhaps unadvisedly) termed "heavy-duty nativism." Many otherwise sympathetic commentators (**Bullinaria, Wells, Elton, Dartnall**) felt this to be a too hasty dismissal of a potentially rich source of recoding functions. With this, however, we have no argument. Our move was rather a strategic one, designed to focus attention on the problematic (but surely inevitable?) residual range of cases in which evolution has not already done our re-coding work for us. We thus accept **Wells's** (see also **Marcus**) suggestion that "evolution's gift of an appropriate set of type-2 problem-relevant recoding biases is exactly what we ought to expect," at least as far as various evolutionarily central learning functions are concerned. But if evolution is to be the *only* source of such re-codings, the lenses of human thought and science must be much weaker and narrower that we had supposed. It seems implausible, to us, to thus limit the space of humanly possible thought (though it surely *has* limits – just not ones directly set by an evolved set of recoding functions). Hence our desire was to explore any other strategies that might be available to us on an ontogenetic *or* cultural-evolutionary time scale.

An interesting suggestion, from **Bullinaria,** is that learning algorithms that build in a few simple and biologically plausible constraints may show improved performance on many problems that would otherwise involve intractable search. Such constraints include assumptions of symmetry between certain weights and the assumption that local information is more likely to matter than distal information. Such fixes and biases constitute, it seems to us, some very plausible ways in which a thrifty nature might subtly bias learning systems so as to promote the successful learning of specific skills in ecologically normal settings (see Karmiloff-

Smith 1992; Clark 1993, Chs. 4 and 5). But once again our primary target lies elsewhere in any residue of cases that must be dealt with by ontogenetic or cultural-evolutionary means.

Similar remarks apply to **Skokowski**'s admonition to look more closely at the architectural inheritance of our biological brains. Of course, as **Dartnall** nicely points out, this is hardly an all-or-nothing matter. Our ontogenetic forays into type-2 space must be primed and rooted in early episodes of thought and learning that exploit quite evolutionarily basic mechanisms for apprehending and responding to our world. Like Dartnall, we envisage a cascade of information-processing activity in which evolution's gifts (**Wells, Skokowski**) and cultural and ontogenetic luck and hard labor come together. Getting straight about their relative contributions and complex interactions is, we think, one of the most pressing tasks facing contemporary cognitive science. The type-1/type-2 distinction is intended as a heuristic tool in the service of just such an endeavor.

If, however, we live in a world in which evolutionarily unanticipated type-2 learning scenarios are regularly encountered, the possibility arises (**Bullinaria, Dominey**) that we may evolve if not fully general, at least multipurpose built-in strategies to support such learning. One plausible contender for such a built-in ploy is whatever onboard machinery supports the process of analogical reasoning. Analogical reason, as noted in our original treatment, provides an open-ended means of re-using achieved recodings so as to view a new problem space through a highly structured lens. Dominey's helpful and illuminating commentary presents a convincing and powerful demonstration of our basic thesis and clearly shows how analogical transfer can at times help to overcome the problem. The central example involved learning sequences not related by surface structure but only by abstract underlying relational structure. Bullinaria shows both the (anticipated) failure of type-1 learning and the success of an augmented model that applies filters (see also **Dartnall**) transferred from other learning experiences. We were especially impressed with Bullinaria's demonstration that the filters for re-use could be selected by a type 1 process of sequence recognition, as this goes some way toward addressing the very real worry (**Wells**) that it is unclear how to choose an achieved recoding for use in a new domain.

A very different set of issues comes to the fore in **Elton's** interesting and provocative comments concerning some important differences between evolutionary and ontogenetic learning scenarios. Elton claims that since problems and solutions can co-evolve it is misleading to think of the issue (over evolutionary time) as one of finding a kind of predetermined target mapping. Instead, he says, it is a matter of finding a kind of behavioral profile that works and then sticking to it. We agree but we do not see that this restatement in any way undermines our project. First, because our principal focus is, as we have said, on individual and perhaps cultural-evolutionary learning. Second, because we already anticipated the role of co-evolution in our original treatment (see our discussion of Newport [1990] in sect. 3). And third, because there is really nothing in our framework that commits us to the view that learning or evolution is anything like passing an exam. In fact our entire argument could be reformulated using Elton's own notion that "creatures stick with [behaviours] that work." Our central notion would become the idea that recoding was only involved in the acquisition of certain "behaviors that work." The rest of our story would remain the same.

R3. Statistics and theories

In pursuing arguments predicated upon the limitations of simple kinds of statistically driven learning, we expose ourselves to the rapid rejoinder that there are simply more things on heaven and earth . . . specifically, what about theory-driven thought, explicit, conscious reflection and the like? Thus **Leiser** reminds us that advanced learning answers to requirements of coherence and coordination and suggests that we should attend more closely to the peculiar dynamics of theory-formation. **Memmi,** likewise, suggests that the presence of rich theoretical contexts and background knowledge deeply inform our advanced searches for recoding biases for relational learning. And **Vinter & Perruchet** (see also sect. R4 below) highlight the role of explicit, conscious reflection in going beyond simple statistical learning.

The general debate here, between theory-based and statistics-based conceptions, is addressed from a connectionist perspective in Clark 1993, Chapter 5. Our general feeling, however, is that the distinction, though clearly important, is easily overplayed. For we still need some account of the *origin* of the theoretical pictures that thus inform subsequent reasoning. And that origin, as far as we can see, can involve only some combination of innate biases and the fruits of an incremental cascade of statistically driven learning.

One crucial link between "mere statistics" and explicit human theorizing is powerfully displayed in **Kurtz**'s very pertinent commentary. Like us, Kurtz hopes to account for complex theory-driven categorization without "quitting the tangible realm of experience and data." This involves, Kurtz suggests, going beyond mere perceptual similarity without losing the solid statistical foundations that are, we believe, the root of all learning. And this in turn involves the recognition and reification of additional *functional* regularities, that is, abstract features that unite disparate instances via the common actions they evoke, the common goals they relate to, and so on. In this way the idea of incrementally constructed, statistically-based feature spaces phases, rather naturally, into the idea of something akin to a theory based take on the various domains of human activity. Our goal is thus not to sever statistical and theory-driven learning but to understand various levels of theoretical understanding as themselves the fruits of an incremental sequence of episodes of type-1 learning, augmented by various tricks involving the re-use of achieved representational resources.

The latter stratagems will often include the uses of analogical reason, and the roles of culture, language, and conscious reflection as stressed by **Memmi, Leiser, Vinter & Perruchet,** and others. But what we should *not* do, we believe, is simply to invoke "theories and background knowledge" as the sufficient answer to the hard question, How is type-2 learning possible at all? For such an invocation is ultimately unexplanatory, trading a problematic chicken for an unexplained egg. Instead, we need to see how such theoretical knowledge can arise from real confrontation with environmental data and how it can be maximally exploited in future learning.

R4. Cognitive psychology

One of the most fruitful and exciting outcomes of the *BBS* process was, for us, the discovery of a quite unexpected wealth of links and connections between our (machine learning based) work and ongoing research in various areas of cognitive psychology, such as implicit learning (**Vinter & Perruchet, Dominey**), analogical reason (**Dominey**), categorization (**Kurtz**), and general research on re-coding (**Haberlandt**). Vinter & Perruchet, in particular, reveal very promising links between some of our claims and work on so-called implicit learning (learning without conscious or linguistic reflection). They agree that the ploys and stratagems we uncover help reveal how very complex problems can be dealt with by elementary, statistics-based processes and note that the course of such learning as predicted by our model is in good accord with experimental research (their own and others) on implicit learning in human subjects. They worry, however (and this ties in with the theory/statistics issues mentioned above) that we fail to address more explicit modes of thought and hence fail to do justice to the full spectrum of human learning.

Our reason for thus stopping short is – as noted above – that we really do believe that in a certain sense there really is no other kind of learning to be had. Such learning, at the very least, lies at the heart of all existing algorithms capable of learning about a domain and not equipped with heavy, task-specific initial biases. We do not deny, of course, that acquired knowledge can be used precisely so as to induce biases that will in effect take the system beyond the domain of visible statistical features of the inputs. Indeed, this is exactly our point: that the only way to go "beyond" the statistics is to use knowledge, itself acquired through earlier instances of type-1 learning (or else innate) so as to re-shape the space for future learning.

Vinter & Perruchet seem to suggest, in addition, that there may be special features of conscious thought and reflection that enable us to do more than simply re-shape the space for learning. Such features would include, for example, the use of "processes which rely on the specific power of conscious thought." Once again, however, we fear a chicken and egg scenario. Our goal is to understand how biological agents can come to wield the knowledge to which such powers may be applied. We do concede, however, that certain aspects of very high level thought look to lie beyond the scope of our treatment. Thus **Vinter & Perruchet** (also **Dartnall**) mention the human ability not just to know a recoding function but to know that we know it. Such knowledge is not of the world so much as of the ways in which we know the world. This certainly does seem like an important ability though the extent to which it figures in daily problem solving is perhaps open to doubt. Whether such top level, meta-reflective capacities merely represent the culmination of a cascade of processes of type-1 learning and re-deployment of achieved representation (as we suspect) or rely on the operation of some wholly different faculty (perhaps tied up with conscious thought) is an important topic for further research. It is, of course, very likely that different neurological structures play a role in type-1 learning and type-2 re-coding (see, e.g., **Dominey**'s comments on the role of the frontostriatal system in type-1 learning). But this is consistent with our claim that the combination of these strategies is effectively all that nature can provide.

Halford's useful and interesting commentary suggests that the type-1/type-2 distinction can help to clarify several issues in the development of children's understanding of mathematics. Halford then makes the important point that not all re-codings are reversible, that is, that it may not be possible to re-create the original data from the recoded data. To avoid this loss of potentially useful information, he suggests a technique that involves representing the input as multiple distinct dimensions that are then processed together. The basic idea sounds interesting, but we found the details of the suggestion elusive. One worry is that, in Halford's own example, the probabilities are based on complete entries in the target mapping. But a complete listing would here constitute a direct recapitulation of the training set – a fact which seems to reduce the technique to the use of a look-up table. In any case, it seems to us that, in this case, one simply cannot have one's code and eat it! In a sense the information-losing properties of the recoding process are crucial since they power the simplification and data compression that in turn lead to the properties of improved search and generalization that the whole process is designed to support. The only real hope in this area, it seems to us, lies in the use of what Karmiloff-Smith (see her 1992, pp. 21–24) once termed conservative redescription – a process in which re-codings are generated but the original representations remain intact and available for use in certain contexts.

In general, then, we were especially pleased to discover these rich links between our themes and treatment and ongoing work in cognitive psychology. One commentator, however (**Ohlsson**), felt that our treatment amounted to a reversion to a discredited behaviorist vision of psychology – one which concerned itself not with the understanding of inner mechanisms but only with patterns of stimulus and response. Here (as with **Elton**) it seems we may have misled by our use of the vocabulary of target mappings, input-output mappings, and so on. But of course we do not wish to claim that no important and contentful inner states mediate between inputs and outputs. Indeed, our whole argument is devoted to displaying the sheer complexity and importance of the search for fruitful inner transformations to be applied to raw input patterns. When Ohlsson says that "the idea that what is learned [by human learners] is an input/output mapping (or a set of stimulus-response connections) was abandoned in the 1950s because people began taking the generativity of human cognition seriously," we are in complete agreement! We are perplexed that Ohlsson sees our paper as in any way disputing this. Our central aim was, in fact, to argue that interesting forms of learning involved not the acquisition of stimulus-response conditions but rather the construction of complex recoding structures (possibly under incremental learning regimes), which would then provide the basis for generative knowledge enabling the learner to go beyond any presented data.

R5. Internal representation

All our talk of inner re-codings raised the hackles of some commentators who seem a little leery of the very idea of internal representation (**Stufflebeam**) or who wanted at least to suggest some possible alternative mechanisms (both internal and external) for achieving the same kinds of result (**Dartnall, Oberlander**). We have already endorsed

the suggestion (**Oberlander, Stufflebeam**) that external structures may sometimes contribute mightily to successful type-2 learning. Stufflebeam seems, in addition, to want to cast doubt on the idea that inner states properly thought of as representation have any role to play in the process at all. We demur, but this is a large and lively debate that we cannot hope to do justice to here – see Clark (1997) for a defense of a kind of modest representationalism consistent with our claims. We would, however, comment that to whatever extent a system creates inner states that effectively reify relational features of the inputs that carry adaptively significant information (as where the chimpanzees learn to match higher order sameness), it is hard to see why we should refuse to label such states internal representations.

Dartnall draws a useful distinction between this (weak) notion of internal representation and one more closely tied to the idea of conscious thought and reflection. Dartnall usefully locates our discussion as part of a larger research program whose goal is to understand the transition between connectionist competence and structured thought. In this context, he suggests that the early stages of our "re-coding" cascade may be more fruitfully conceived in terms of a sequence of increasingly powerful ways of accessing the knowledge rather than in terms of re-codings of the knowledge itself. This is an interesting idea and one that seems to invite a slightly different perspective on Karmiloff-Smith's notion of representational re-description – a notion that provided much of the motivation and inspiration for the present treatment [see multiple book review of Karmiloff-Smith's *Beyond Modularity BBS* 17(4) 1994]. We are not yet convinced, however, that this distinction (between changing access and changing knowledge) is deep and well-defined. But it certainly suggests some directions for future research, perhaps using some concrete computational models to refine our currently hazy intuitions.

R6. Technicalia

A number of commentators made useful technical suggestions concerning the description of the type-1/type-2 distinction, and raised important questions about the relations between the parity considerations and generalization and the example of the Elman net. Thus **Damper** (see also **Gaskell**) worries that holding back even a single pattern on the classical (2 variable, XOR) parity problem simply makes the problem insoluble (the machine would need to read our minds to know the intended function) as the learning algorithm lacks sufficient data. He concludes that it must be wrong to link parity learning to issues about generalization. But let us step back a little here and review our strategy in more detail. In the paper we suggested that backpropagation learning does not provide a ready-made solution to the problem of type-2 scenarios and backed this up with a demonstration that backpropagation reliably fails on some forms of parity-generalization problem. The slight novelty in this was indeed the utilization of parity as a generalization problem. Where parity problems have been used in machine learning, they have typically been presented as memory tasks, that is, learning methods have been required to acquire complete mappings. One of the justifications put forward for this approach is the idea that parity constitutes an "unfair" generalization problem. Damper's commentary is valuable because is shows how muddled the thinking behind this judgment can be.

Damper implies that parity cannot be a generalization problem because parity mappings exhibit neutral statistics, that is, chance-level output probabilities. This observation was a fundamental component in our own presentation. But it demonstrates not that parity problems are ungeneralizable but merely that they cannot be generalized on the basis of statistical effects.

In coming to terms with the idea of parity generalization, it is useful to turn attention away from the familiar XOR case towards higher-order cases. In 4-bit parity there are 16 cases. The removal of a single case leaves 15 cases as a basis for generalization. Somehow this does not seem quite so unreasonable. It may also be helpful to consider the two-spirals problem in which the learning algorithm must learn to correctly assign a 2-d point to one of two interlocking spirals in an image. The problem is parity-like since nearest-neighbors in the input space always have opposite classifications. And indeed, the statistics of a typical training set are usually nearly neutral. And yet, as far as we are aware, this problem has never been treated as anything other than a generalization problem.

Despite **Damper**'s objections to our use of parity generalization problems, he has no difficulty with our central thesis that learning problems which require recoding present a special and cognitively important case.

Turning to the type-1/type-2 distinction itself, **Chater** argues that the distinction is simply ill-defined and hence will confuse rather than clarify matters. Notice, however, that we took some care, in section 1 of the paper, to stress that "there is no obvious operational definition for the class of type-2 problems." However, by proceeding on the basis that such a definition exists, and indeed that our paper was supposed to provide it, **Chater** has arrived at a number of interesting though strictly speaking irrelevant observations. His approach involves taking our analysis of the ways in which supervisory feedback can provide justifications for assignments of particular probabilities to particular outputs as a formal definition of a problem class. He shows that this soon leads to nonsensical results and enables dubious maneuvers such as the adding of "dummy variables" so as to change the "formal" characterisation of a particular problem.

These arguments may be of interest to the computational learning theorist. However, they completely miss the point of our paper and in some cases actually mislead. For example, **Chater** contrives to show that if we try to give identity-function learning a classification using his "formalization" of our framework, certain ambiguities result. However, this conceals that fact that identity-function learning actually has a rather natural characterization as a type-2 operation within our framework.

Assume that the input for the learner is based on two variables – one representing the input to the identity function and the other representing the output – and that the target output for the learner is a value which shows whether the input forms a valid application of the identity function (i.e., whether the two input values are the same). In this scenario the learner's guessing of a particular output cannot be justified on the basis of observed frequencies in the training data since every input value is unique. However, were we to recode the learner inputs by applying an identity recognition function to them, we would produce a

recoding of the problem which could be solved in exactly that "statistical" way. Thus identity function learning is naturally characterized in our framework as requiring recoding and hence is type-2.

Golden, in his gracious and intriguing commentary, offers an amendment to our account of type-1 learning. He suggests that the category can be broken down into subcases using a "parametric, model-based approach" and that this may help avoid some potential problems. Alas, we are not sufficiently familiar with the background material for this proposal to properly judge its value or necessity. We agree, however, that there may well exist other (perhaps more elegant) ways of picking apart the kinds of cases we wish to distinguish, and we look forward to Golden (forthcoming) to learn more about the methods he has in mind.

Finally, some questions were raised concerning the role of the Elman network in our discussion. Thus **Marcus** argues that the Elman network (Elman 1993) fails to illustrate our central point as it does not, after all, learn about higher order regularities in the data set. The evidence that Marcus relies on, however, concerns only the failure of such a net to generalize an abstract structure (an X is an X) when presented with a case involving a totally novel "filler" (a dax is an . . .). It may be that such extensions simply require more than the kind of grammatical knowledge that the data set makes available. In any case, it does not follow from this kind of failure that the original network does not acquire grammatical knowledge that is higher order in the sense we require. For the successful network did indeed proceed by first identifying lower level grammatical features and then going on to learn about regularities involving relations between these lower level features. In fact, this is exactly what the incremental batching/staged memory manipulations were designed to encourage. **Gaskell** seems to worry that the need for such manipulations renders the network unsuitable for our purposes. We do not see why: our point here is simply that the desired performance is achieved only by the prior isolation (by whatever means) of a "building block" set of regularities which then mediate between the raw data and the target mapping so as to shrink the search space to a manageable size. The network thus trades prior learning against subsequent search.

Taken together, the various commentaries have done much to advance (and where necessary, to unsettle) our thinking about the nature of learning and the canny ways in which biological cognizers may trade achieved representation against potentially infinite computational search. Our treatment, we readily concede, can easily appear either trite or wildly speculative. Few disagree with the central tenet (re-coding matters!). Few will concede our central claim (that the maximal exploitation of the fruits of simple learning or chance penetrations into type-2 space is the best nature can provide). We hope, however, to have fueled the fires of debate. And we thank all the commentators for their thoughtful and genuinely constructive responses. We have learnt a lot, and we look forward to trading it against our future search in computational space.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Agre, P. E. (1993) The symbolic worldview: Reply to Vera and Simon. *Cognitive Science* 17(1):61–69. [RSS]

Agre, P. E. (1995) Computation and embodied agency. *Informatica* 19(4):527–35. [RSS]

Anderson, J. A. (1995) *An introduction to neural networks*. MIT Press. [KH]

Anderson, J. R. (1983) *The architecture of cognition*. Harvard University Press. [KH]

Armstrong, S. L., Gleitman, L. R. & Gleitman, H. (1983) What some concepts might not be. *Cognition* 13:263–308. [KK]

Ausubel, D. (1963) *The psychology of meaningful verbal learning.* Grune & Stratton. [SO]

Barkow, J. H., Cosmides, L. & Tooby, J. (1992) *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press. [AW]

Baum, E. B. & Haussler, D. (1989) What size net gives valid generalization? *Neural Computation* 1:151–60. [JAB]

Baum, E. B. & Wilczek, F. (1988) Supervised learning of probability distributions by neural networks. In: *Neural information processing systems*, ed. D. Z. Anderson. American Institute of Physics. [RID]

Baxter, J. (1995) Learning internal representations. In: *Proceedings of the Eighteenth International Conference on Computational Learning Theory*. ACM Press. [NS]

Becker, S. & Cun, Y. (1988) *Improving the convergence of back-propagation learning with second-order methods [CRG-TR-88–5]*. University of Toronto Connectionist Research Group. [aAC]

Berkeley, I., Dawson, M., Medler, D., Schopflocher, D. & Hornsby, L. (1995) Density plots of hidden unit activations reveal interpretable bands. *Connection Science* 7(2):167–86. [ISNB]

Brooks, R. (1991) Intelligence without representation. *Artificial Intelligence* 47:139–59. [RSS]

Bullinaria, J. A. (1994) Modeling reading, spelling and past tense learning with artificial neural networks. *Brain and Language* (in press). [JAB]

Campbell, D. (1974) Evolutionary epistemology. In: *The philosophy of Karl Popper*, ed. P. Schillp. Open Court. [aAC]

Carey, S. & Spelke, E. (1994) Domain-specific knowledge and conceptual change. In: *Mapping the mind: Domain specificity in cognition and culture*, ed. L. A. Hirschfeld & S. A. Gelman. Cambridge University Press. [DL]

Caruana, R. A. (1994) Multitask connectionist learning. In: *Proceedings of the 1993 Connectionist Summer School.* Erlbaum. [NS]

Chi, M. T. H. (1992) Conceptual change within and across ontological categories. In: *Cognitive models of science*, ed. R. N. Giere. University of Minnesota Press. [DL]

Chomsky, N. (1959) Review of Skinner's "Verbal behavior." *Language* 35:26–58. [SO]

Chomsky, N. & Miller, G. (1963) Introduction to the formal analysis of natural language. In: *Handbook of mathematical psychology*, vol. 2, ed. R. D. Luce, R. R. Bush & E. Galanter. Academic Press. [aAC]

Churchland, P. M. (1995) *The engine of reason, the seat of the soul.* MIT Press. [aAC]

Churchland, P. & Sejnowski, T. (1992) *The computational brain.* MIT/Bradford. [aAC]

Clark, A. (1989) *Microcognition.* MIT Press. [rAC]

(1993) *Associative engines: Connectionism, concepts and representational change.* MIT/Bradford. [arAC]

(1997) *Being there: Putting brain, body and the world together again.* MIT Press. [RSS, rAC]

Clark, A. & Karmiloff-Smith, A. (1993) The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language* 8:487–519. [aAC]

Clark, A. & Toribio, J. (1994) Doing without representing? *Synthese* 101:401–31. [RSS]

Cloete, I. & Ludik, J. (1993) Increased complexity training. In: *Proceedings: New trends in neural computation*, ed. J. Mira, J. Cabestony, & A. Prieto. Lecture notes in computer science, no. 686. Springer-Verlag. [NS]

Dawson, M. & Schopflocher, D. (1992) Modifying the generalized Delta Rule to train networks of nonmonotonic processors for pattern classification. *Connection Science* 4(1):19–31. [ISNB]

Dennett, D. (1991) *Conciousness explained.* Little, Brown. [aAC]

(1994) Labeling and Learning: Commentary on Clark and Karmiloff-Smith. *Mind and language* 8:544–548. [aAC]

Dienes, Z. & Berry, D. (in press) How implicit is implicit learning? *Psychonomic Bulletin and Review.* [AV]

Dietterich, T., London, B., Clarkson, K. & Dromey, G. (1982) Learning and inductive inference. In: *The handbook of artificial intelligence*, vol. 3, ed. P. Cohen & E. Feigenbaum. Morgan Kaufmann. [aAC]

Dominey, P. F. (1995) Complex sensory-motor sequence learning based on recurrent state-representation and reinforcement learning. *Biological Cybernetics* 73:265–74. [PFD]

Dominey, P. F. (in press) An anatomically structured sensory-motor sequence

learning system displays some general linguistic capacities. *Brain and Language.* [PFD]
Dominey, P. F., Arbib, M. A. & Joseph, J. P. (1995) A model of cortico-striatal plasticity for learning oculomotor associations and sequences. *Journal of Cognition and Neuroscience* 7(3):311–36. [PFD]
Dominey, P. F., Ventre-Dominey, J., Broussolle, E. & Jeannerod, M. (1995) Analogical transfer in sequence learning: Human and neural-network models of fronto-striatal function. *Annals of the New York Academy of Sciences* 769:369–73. [PFD]
Dominey, P. F., Ventre-Dominey, J., Broussolle, E. & Jeannerod, M. (in press) Analogical transfer is effective in a serial reaction time task in Parkinson's disease: Evidence for a dissociable sequence learning mechanism. *Neuropsychologia.* [PFD]
Dreyfus, H. L. (1992) *What computers still can't do: A critique of artificial reason,* rev. ed. MIT Press. [RSS]
Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition* 48:71–99. [arAC, JAB, RID, MGG, GFM, DM, KH, NS, PS, TD]
Eysenck, M. W. & Keane, M. T. (1995) *Cognitive psychology: A student's handbook.* Erlbaum. [SO]
Fahlman, S. & Lebiere, C. (1990) *The cascade-correlation learning architecture [CMU-CS-90-100].* Carnegie-Mellon University, School of Computer Science, Pittsburgh, PA 15213. [aAC]
Fahlman, S. E. & Lebiere, C. (1990) The cascade correlation learning architecture. In: *Advances in neural information processing systems 2,* ed. D. Touretzky. Morgan Kaufman. [NS]
Fodor, J. A. (1975) *The language of thought.* Harvard University Press. [RSS]
(1995) The folly of simulation. In: *Speaking minds: Interviews with twenty eminent cognitive scientists,* ed. P. Baumgartner & S. Payr. Princeton University Press. [RSS]
Funt, B. V. (1980) Problem solving with diagrammatic representations. *Artificial Intelligence* 13:210–30. [JO]
Gagne, R. M. (1962) The acquisition of knowledge. *Psychological Review* 69:355–65. [SO]
Gallant, S. I. (1994) *Neural network learning and expert systems.* MIT Press. [KH]
Gallistel, R. (1994) Interview. *Journal of Cognitive Neuroscience* 6(2):174–79. [aAC]
Gentner, D., Rattermann, M. J. & Forbus, K. D. (1993) The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology* 25:524–75. [DL]
Gick, M. & Holyoak, K. J. (1980) Analogical problem solving. *Cognitive Psychology* 12:306–55. [NS]
(1983) Schema induction and analogical transfer. *Cognitive Psychology* 15:1–38. [NS]
Gold, R. (1987) *The description of cognitive development: Three Piagetian themes.* Oxford University Press. [DL]
Goldberg, D. (1989) *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley. [aAC]
Golden, R. M. (1988) A unified framework for connectionist systems. *Biological Cybernetics* 59:109–20. [RMG]
(forthcoming) *Mathematical methods for neural network analysis and design.* MIT Press. [RMG, rAC]
Goldstone, R. L. (1994) The role of similarity in categorization: Providing a groundwork. *Cognition* 52:125–57. [KK]
Gomez, R. L. & Schaveneveldt, R. W. (1994) What is learned from artificial grammars? Transfer tests of simple association. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(2):396–410. [PFD]
Goswami, U. (1991) Analogical reasoning: What develops? A review of research and theory. *Child Development* 62:1–22. [DL]
Haberlandt, K. F. (1990) Expose hidden assumptions in network theory. *Behavioral and Brain Sciences* 13:495–96. [KH]
Haberlandt, K. F., Graesser, A. C., Schneider, N. J. & Kiely, J. (1986) Effects of task and new arguments on word reading times. *Journal of Memory and Language* 25:314–22. [KH]
Halford, G. S. (1993) *Children's understanding: The development of mental models.* Erlbaum. [GSH]
Halford, G. S., Wilson, W. H. & Phillips, S. (submitted) Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. [GSH]
Hare, M., Elman, J. & Daugherty, K. (1995) Default generalization in connectionist networks. *Language and Cognitive Processes* 10:601–30. [GFM]
Hertz, J., Krogh, A. & Palmer, R. G. (1991) *Introduction to the theory of neural computation.* Addison-Wesley. [RID]
Hinton, G. & Sejnowski, T. (1986) Learning and relearning in Boltzmann machines. In: *Parallel distributed processing: Exploration of the microstructures of cognition,* vols. 1 & 2, ed. D. Rumelhart, J. McClelland & the PDP Research Group. MIT Press. [aAC, NC]
Hutchins, E. (1995) *Cognition in the wild.* MIT Press. [rAC]
Jackson, G. M., Jackson, S. R., Harrison, J., Henderson, L. & Kennard, C. (1995) Serial reaction time learning and Parkinson's disease: Evidence for a procedural learning deficit. *Neuropsychologia* 33(5):577–93. [PFD]
Jacobs, R., Jordan, M. & Barto, A. (1991a) Task decomposition through competition in a modular connectionist architecture: The what and where visual tasks. *Cognitive Science* 15:219–50. [aAC]
Jacobs, R., Jordan, M., Nowlan, S. & Hinton, G. (1991b) Adaptive mixtures of local experts. *Neural Computation* 3:79–87. [aAC]
Jarvella, R. J. (1979) Immediate memory and discourse processing. In: *The psychology of learning and motivation: Advances in research and theory,* ed. G. H. Bower. Academic Press. [KH]
Karmiloff-Smith, A. (1979) *A functional approach to child language.* Cambridge University Press. [aAC]
(1992a) Nature, nurture and PDP: Preposterous development postulates? *Connection Science* [special issue on philosophical issues in connectionist modelling] 4(3/4):253–70. [arAC, TD, NS]
(1992) *Beyond modularity: A developmental perspective on cognitive science.* MIT Press. [DM]
Karmiloff-Smith, A. & Clark, A. (1993) What's special about the development of the human mind/brain? *Mind and Language* 8(4):569–81. [aAC]
Keil, F. C. (1994) The birth and nurturance of concepts by domains. In: *Mapping the mind: Domain specificity in cognition and culture,* ed. L. A. Hirschfeld & S. A. Gelman. Cambridge University Press. [DL]
Kintsch, W. (1988) The use of knowledge in discourse processing: A construction-integration model. *Psychological Review* 95:163–82. [KH]
Kirsh, D. (1991) When is information explicitly represented? In: *Information, thought and content,* ed. P. Hanson. UBC Press. [aAC]
Kirsh, D. (1995) The intelligent use of space. *Artificial Intelligence* 72:1–52. [aAC]
Kirsh, D. & Maglio, P. (1994) On distinguishing epistemic from pragmatic action. *Cognitive Science* 18:513–19. [aAC]
Krogh, A. & Hertz, J. A. (1992) A simple weight decay can improve generalization. In: *Advances in neural information processing systems,* ed. J. E. Moody, S. J. Hanson & R. P. Lippman. Morgan Kauffman. [JAB]
Kurtz, K. J. (1996) *Category-based similarity.* Poster presented at the Eighteenth Annual Conference of the Cognitive Science Society, San Diego, CA. [KK]
Larkin, J. H. & Simon, H. A. (1987) Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science* 11:65–100. [JO]
LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard,W. & Jackal, L. (1989) Back propagation applied to handwritten zipcode recognition. *Neural Computation* 1:541–51. [aAC]
Leiser, D. (1996) Computational emergence and developmental emergence: A sobering survey. *Intellectica* 22:203–220. [DM]
Lewicki, P., Hill, T. & Bizot, E. (1988) Acquisition of procedural knowledge about a pattern of stimuli that can not be articulated. *Cognitive Psychology* 20:24–37. [AV]
Lindsay, R. K. (1988) Images and inference. *Cognition* 29:229–50. [JO]
MacKay, D. J. C. (1992a) Bayesian interpolation. *Neural Computation* 4:415–47. [JAB]
(1992b) A practical Bayesian framework for backpropagation networks. *Neural Computation* 4:448–72. [NC]
Marcus, G. F. (submitted a) Rethinking eliminative connectionism. [GFM]
(submitted b) What does it take to get a connectionist model to generalize a low frequency default? [GFM]
Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R. & Pinker, S. (1995) German inflection: The exception that proves the rule. *Cognitive Psychology* 29:186–256. [GFM]
Marr, D. (1982) *Vision.* Freeman. [aAC]
McClelland, J. L. & Rumelhart, D. E. (1988) *Exploration in parallel distributed processing: A handbook of models, programs, and exercises.* MIT Press/Bradford. [RID]
Miller, G. A. (1956) The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63:81–93. [SO, KH]
Minsky, M. & Papert, S. (1988) *Perceptrons: An introduction to computational geometry,* 3rd ed., expanded. MIT Press. [aAC, ISNB]
Murphy, G. L. & Medin, D. L. (1985) The role of theories in conceptual coherence. *Psychological Review* 92:289–316. [KK]
Newport, E. (1990) Maturational constraints on language learning. *Cognitive Science* 14:11–28. [arAC, TD]
Nissen, M. J. & Bullemer, P. (1987) Attentional requirements of learning: Evidence from performance measures. *Cognition Psychology* 19:1–32. [PFD]
Nowlan, S. N. & Hinton, G. E. (1992) Simplifying neural networks by soft weight sharing. *Neural Computation* 4:473–93. [JAB]

Ohlsson, S. (1996) Learning from performance errors. *Psychological Review* 2:241–62. [SO}

Ohlsson, S. & Rees, E. (1991) The function of conceptual understanding in the learning of arithmetic procedures. *Cognition and Instruction* 8:103–79. [SO]

Perruchet, P. & Gallego, J. (in press) A subjective unit formation account of implicit learning. In: *How implicit is implicit learning?* ed. D. Berry. Oxford University Press. [AV]

Perruchet, P., Gallego, J. & Savy, I. (1990) A critical reappraisal of the evidence for unconscious abstraction of deterministic rules in complex experimental situations. *Cognitive Psychology* 22:493–516. [AV]

Piaget, J. (1971) *Biology and knowledge.* Chicago University Press. [DM]

(1974) *Adaptation vitale et psychologie de l'intelligence.* Hermann. [DL]

Port, R. & Van Gelder, T. (1995) *Mind as motion: Explorations in the dynamics of cognition.* MIT Press. [RSS]

Quinlan, J. (1986) Induction of decision trees. *Machine Learning* 1:81–106. [aAC]

(1993) *C4.5: Programs for machine learning.* Morgan Kaufmann. [aAC]

Reber, A. S. (1993) *Implicit learning and tacit knowledge: An essay on the cognitive unconscious.* Oxford University Press. [KH]

Reeves, L. M. & Weisbert, R. W. (1994) The role of content and abstract information in analogical transfer. *Psychological Bulletin* 115:381–400. [DL]

Rendell, L. (1989) A study of empirical learning for an involved problem. In: *Proceedings of the Eleventh Joint Conference on Artificial Intelligence.* Morgan Kaufmann. [aAC]

Rescorla, R. A. & Wagner, A. R. (1972) A theory of Pavlovian conditioning: The effectiveness of reinforcement and nonreinforcement. In: *Classical conditioning 2: Current research and theory,* ed. A. H. Black & W. F. Prokasy. Appleton-Century-Crofts. 64–69. [RID]

Richard, M. E. & Lippman, R. P. (1991) Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation* 3:461–83. [NC]

Rozin, P. (1976) The evolution of intelligence and access to the cognitive unconscious. In: *Progress in psychobiology and physiological psychology,* ed. J. Sprague & A. Epstein. Academic Press. 245–80. [AW]

Rudy, J. W. (1991) Elemental and configural associations, the hippocampus and development. *Developmental Psychobiology* 24(4):221–36. [GSH]

Rumelhart, D. E., Durbin, R., Golden, R. & Chauvin, Y. (1995) Backpropagation: The basic theory. In: *Backpropagation,* ed. D. E. Rumelhart & Y. Chauvin. Cambridge, MA: MIT Press. Bradford Books. [KK]

Rumelhart, D. E., Hinton, G. & Williams, R. (1986) Learning representations by backpropagating errors. *Nature* 323:533–36. [aAC, ISNB, KH, KK]

Schultz, T. R., Schmidt, W. C., Buckingham, D. & Mareshal, D. (1995) Modeling cognitive development with a generative connectionist algorithm. In: *Developing cognitive competence: New approaches to process modeling,* ed. T. Simon & G. Halford. Erlbaum. [DL]

Searle, J. R. (1990) Is the brain a digital computer? *APA Proceedings* 64(3):21–37. [RSS]

Sejnowski, T. & Rosenberg, C. (1987) Parallel networks that learn to pronounce English text. *Complex Systems* 1:145–68. [aAC]

Sejnowski, T. J. & Rosenberg, C. R. (1986) *NETtalk: A parallel network that learns to read aloud (Report No. JHU/EECS - 86/01).* Johns Hopkins University. [KH]

Stenning, K. & Oberlander, J. (1995) A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science* 19:97–140. [JO]

Stufflebeam, R. S. (1995) Representations, explanation, and PDP: Is representation talk really necessary? *Informatica* 19(4):599–613. [RSS]

(1997) *Whither internal representations? In defense of antirepresentationalism and other heresies.* Doctoral dissertation, Washington University, St. Louis, MO. [RSS]

Sun, R. & Bookman, L. (1994) *Computational architectures integrating neural and symbolic processes.* Kluwer. [DM]

Sutton, R. S. & Barto, A. G. (1981) Towards a modern theory of adaptive networks: Expectation and prediction. *Psychological Review* 88:135–70. [RID]

Szilas, N. & Ronco, E. (1995) Action for learning in nonsymbolic systems. In: *Proceedings of the European Conference on Cognitive Science.* INRIA. [NS]

Tetewsky, S., Shultz, T. & Buckingham, D. (1993) Reducing retroactive interference in connectionist models of recognition memory. Canadian Society for Brain, Behaviour, and Cognitive Science. Third Annual Meeting. [NS]

Tetewsky, S. J., Shultz, T. R. & Takane, Y. (1995) Training regimens and function compatibility: Implications for understanding the effects of knowledge on concept learning. In: *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society.* Erlbaum. 304–309. [NS]

Thagard, P. (1989) Explanatory coherence. *Behavioral and Brain Sciences* 12:435–502. [DL]

Thagard, P., Holyoak, K. J., Nelson, G. & Gochfeld, D. (1990) Analogical retrieval by constraint satisfaction. *Artificial Intelligence* 46:259–310. [PFD]

Thompson, R., Oden, D. & Boyson, S. (in press) Language-naive chimpanzees (pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes.* [rAC]

Utgoff, P. (1986) *Machine learning of inductive bias, vol. 15 (Kluwer International Series in Engineering and Computer Science).* Kluwer. [aAC]

Valiant, L. G. (1984) A theory of the learnable. *Communications of the ACM* 27:1134–42. [NC]

Van Gelder, T. (1995) What might cognition be if not computation? *Journal of Philosophy* 92(7):345–81. [RSS]

Vinter, A. & Perruchet, P. (1994) Is there an implicit level of representation? *Behavioral and Brain Sciences* 17:730–31. [AV]

Vosniadou, S. & Brewer, W. F. (1987) Theories of knowledge restructuring in development. *Review of Educational Research* 57:51–67. [DL]

White, H. (1989) Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association* 84:1003–13. [RMG]

Wilson, A. & Hendler, J. (1993) Linking symbolic and subsymbolic computing. *Connection Science* 5:3–4. [DM]

Wisniewski, E. J. & Medin, D. L. (1994) On the interaction of theory and data in concept learning. *Cognitive Science* 18:221–81. [KK]