Does generalization in infant learning implicate abstract algebra-like rules?

James L. McClelland and David C. Plaut

Two recent papers[1,2] suggest that infants well under a year old can learn from exposure to relatively short samples of language-like sequences of syllables. The first of these two papers, by Saffran and colleagues[1], suggested that infants were sensitive to the statistical structure of syllable sequences that they heard, in that they tended to listen longer to syllable sequences that were less common in their brief experience. An interesting exchange of commentaries ensued[2], centering mostly on the issue of whether it is surprising that infants are sensitive to the statistical structure of experience. No one has admitted to being surprised; what the debate has been about is whether there are any theoretical positions under which anyone should have been surprised. We think enough has been said about that.

The findings reported in the second of these papers, by Marcus and colleagues[3], have raised a rather different debate, this one centering on whether or not infants learn something over and above the statistical structure of the language-like sequences that they hear. Specifically, the authors found that, following exposure to syllable sequences obeying a particular pattern, seven-month-old infants tended to listen longer to syllable sequences that violated the pattern, even when the new sequences were composed of novel syllables. The authors suggest that their findings cannot be accounted for by any learning method that relies on statistical information. Rather, they claim, the results implicate the learning and use of abstract 'algebra-like' rules.

Before we review the findings in more detail, we will lay our cards on the table: We don't think the findings of Marcus and colleagues really cut any ice at all regarding the possible existence of abstract rule learning in infants. We will support this position by describing several ways in which the data might be seen as compatible with the extraction of statistical information rather than abstract rules. We do think the question of the extent of infants' generalization ability is very interesting, because it tells us something about what they bring to learning situations at a relatively early age. Indeed, it seems very possible to us that seven-month-old infants possess mechanisms that provide powerful support for generalization. But we don't really see how experiments of this general sort can tell us whether they use rules *per se*; the powerful mechanisms might simply be ones that help statistical learning procedures generalize in powerful ways. Furthermore, these mechanisms might themselves be learned.

J.L. McClelland and D.C. Plaut are at the Center for the Neural Basis of Cognition, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213-2683, USA.

tel: +1 412 268 4000
fax: +1 412 268 5060
e-mail: jlm@cnbc.cmu.edu, plaut@cmu.edu

The case for rules

Marcus *et al.* based their argument on the findings of three experiments. We will focus on the third experiment, as it provided the strongest test of the authors' hypothesis. During the training phase of the experiment, each infant was exposed to several examples of sequences of three syllables conforming to a simple general pattern: either AAB or ABB. That is, half of the infants heard a series of sequences like '*de-de-li*', '*wi-wi-di*', etc., while the other half heard sequences like '*de-li-li*', '*wi-di-di*', etc. In the subsequent test phase, the infants' listening preferences were evaluated for both AAB and ABB sequences. Importantly, the test sequences were composed of new syllables that had not been presented during the training phase (e.g. '*ba-ba-po*'). Moreover, to preclude statistical learning at the phonetic feature level, the new sequences varied only in phonetic features (e.g. consonant voicing) that had been held constant across training sequences; the training phase would thus provide no basis for preferring any sequence of test syllables over another. Marcus *et al.* found that, despite the fact that both AAB and ABB test sequences contained repetition and that variation in phonetic features during training was uninformative at test, infants nonetheless tended to listen less to test sequences that obeyed the pattern to which they had been previously exposed. Based on these results, the authors rejected several varieties of 'statistical learning mechanisms' and instead suggested:

> We propose that a system that could account for our results is one in which infants extract abstract algebra-like rules that represent relationships between placeholders (variables) such as 'the first item X is the same as the third item Y' or more generally that 'item I is the same as item J'. (Ref. 2, p. 79.)

Why the case is not convincing

We certainly agree that a system that extracts abstract algebra-like rules could account for the results. On the other hand, it might be preferable to avoid postulating a separate rule-learning system if statistical learning mechanisms, which everyone seems to agree are involved in early language acquisition, could be shown to be sufficient. In evaluating this possibility, it is important to be clear on a key point that is often misunderstood in discussions of what statistical learning mechanisms can and cannot use as a basis for generalization. In making this point, we will focus on neural networks as one form of such a mechanism, although we believe our comments apply to a much broader class of statistical approaches.

The point we wish to make relates to the fact that generalization in neural networks depends on overlap of representations – that is, the patterns of activity used in the network – to represent items experienced during training and test. For prior learning to generalize to a new stimulus, the representation of the new stimulus must overlap with – that is, activate some units in common with – the representation of the stimuli on which learning is based. This is because learning occurs by the adjustment of connection weights between specific units in a network, and so a new input must activate some of the same units whose weights were influenced by prior experience to benefit from that experience.

We can now turn to our key point, which is that the characteristic of neural networks just described has been misconstrued as implying that the relevant overlap must be present in the input itself. For example, Putnam claims:

1364-6613/99/$ – see front matter PII: S1364-6613(99)01320-0

> … what [the connectionist] algorithm does is find correlations between sets of variables that it is given. Things that cannot be expressed in terms of correlations between given variables cannot be found by the algorithm. That is how Steven Pinker [Ref. 5] at MIT was able to show that you cannot claim that neural nets learn the past tense in English. In principle, they cannot, because there are features in the past tense for example, the division of verbs into classes that form their past tense differently that cannot be expressed simply as correlations between variables given to the machine as input. (Ref. 4, p. 185.)

As this quotation suggests, this idea that generalization in neural networks must depend on given variables may have played into several authors' rejection of neural network models. If so, the rejection might have been premature, as it neglects the following crucial fact:

The relevant overlap of representations required for generalization in a neural network or other statistical learning procedure need not be present directly in the 'raw input' but can arise over internal representations that are subject to learning.

It is important to remember that infants are constantly bombarded by auditory and linguistic information that might contribute to the formation of such representations, and indeed many models have been developed in which learning through adjustment of connection weights accounts for the formation of a variety of aspects of brain representations, including receptive field properties of neurons in visual cortex and the organization of topographic maps[6–10].

Alternatives to using rules for generalization

With this background, we submit that a range of different proposals might be made about what sorts of map-pings from raw input might give rise to representational overlap that would support generalization of the kind observed in the Marcus *et al.* experiments.

The first possibility is perhaps the least interesting: that both the training and test stimuli vary along one or more acoustic dimensions that are not captured by the phonetic features that Marcus *et al.* considered. Marcus and colleagues were correct to realize that the argument for abstract rules would be undermined if there were dimensions of variation among syllables in the training sequences that also varied among the test stimuli, because then the results could be due to simple distributional learning along these dimensions. The problem is that, even if one assumes that seven-month-old infants have a level of representation corresponding to the specific phonetic feature set envisioned by the experimenters (and this is a big 'if'), there is no compelling reason to think that this would be the *only* level of representation that could influence the infants' listening preferences. For example, suppose that, as seems likely, the training syllables and the test syllables both vary along an acoustic dimension like loudness. If so, an AAB sequence in both the training and test phases might be loud–soft–soft or soft–loud–loud, and simple statistical learning would be sufficient to pick up on repetition along this dimension. Loudness is, of course, just one possibility; other possibly relevant acoustic dimensions include the frequencies, amplitudes, and transition rates of various formants in the consonants or vowels contained in the syllables used. In essence, this first proposal amounts to a suggestion that representational overlap rather close to the 'raw input' might be sufficient to account for the generalization demonstrated in the Marcus *et al.* experiments. We simply don't know whether this is a viable possibility since the authors report no attempts to assess or control for any acoustic properties of their stimuli.

The remaining possibilities we will consider are more interesting in that they accept, to different degrees and in different ways, that the performance of the infants at test is based on some form of abstraction. Where they differ from the account offered by Marcus and colleagues is that the process of abstraction reflects a recoding of the input, which then makes abstract information available for further computations (of whatever form). If the uninteresting possibility discussed above can be ruled out, we think the evidence is interesting in suggesting that infants are capable of some form of abstraction; but we would suggest that the data do not cut either way with respect to whether the learning that applies to these abstractions involves statistical computations or rule induction.

Perhaps the simplest form of abstraction would be a case in which representational overlap is not present in the input but is introduced by the application of some form of normalization or relative encoding. Pursuing the loudness example further, it might be the case that the loudness in the training stimuli varies in a different range from the loudness in the test stimuli. If different values of loudness were represented by different units in a network, there might be no overlap. However, if loudness is encoded relative to other adjacent stimuli (and we know such relative coding is used for brightness and many other visual qualities) then patterns such as 'very loud–loud–loud' for a training sequence, and 'medium–soft–soft' for a test sequence, would both map onto the same values: louder–softer–softer. Given a normalization process operating over stimuli, any movement of the relative value of any auditory dimension becomes a potential basis for generalization from the training to the test stimuli in the Marcus *et al.* experiments.

Another possibility, involving a stronger form of abstraction, would be if the infant's perceptual system encoded whether a perceptual input is the same or different from other items in the immediate context. (Indeed, the application of the preferential-looking method to test infants' discrimination abilities is predicated on sensitivity to repetition/novelty.) Based on the quote printed above, Marcus and colleagues appear to assume that infants encode instances of sameness among all combinations of positions within a string of syllables. Without such information, it is unclear how infants might discover the rules attributed to them. If sameness is assumed to be available as input to a rule-learning mechanism, we see no reason why it should not also be available to the statistical learning mechanism. In this case (considering only the last two syllables), AAB patterns end with same–different whereas ABB patterns end with different–same, and statistical learning would again be sufficient for generalizing from training to test sequences. In fact, in a response to the Marcus *et al.* article, Seidenberg and Elman[11] refer to a neural network simulation that learned to detect patterns of repetition within syllable strings in a way that allowed later learning to generalize to sequences composed of novel syllables constructed to be analogous to the Marcus *et al.* stimuli. One may quibble with the particulars of the reported simulation, but the general point remains that an encoding of sameness versus difference may be available as a basis for various forms of learning (statistical or otherwise) in seven-month-old infants.

The final possibility we will consider is the suggestion, arising from the work of Dienes, Altmann and Gao[12], that learning during the *test* phase contributes to inducing overlap in the internal representations for the test and training sequences, even if there is no overlap in their input representations. This can arise through learning to map the test syllables onto the same internal units that encode the elements of the trained sequences. In these studies, the network is trained simply to predict each upcoming syllable within sequences, without regard to what pattern the sequence obeys. Dienes and colleagues have used this method successfully to account for transfer of knowledge of sequential structure based on one set of elements to an entirely new set of elements. A forthcoming paper[13] demonstrates clearly that transfer of such knowledge without any input overlap is quite possible, without the use of abstract,

algebraic rules. In their simulations (as in the Marcus *et al.* experiment) only half the test sequences have the same structure as the sequences used in training. Nonetheless, the learning process quickly induces similarity among the novel and familiar syllables. As a result, sequences made from the new elements cannot help but tap into the knowledge the system has built up about the sequential structure present in the trained sequences, thereby producing generalization.

In summary, we have described a number of possible ways in which the type of generalization exhibited by infants in the Marcus *et al.* experiments might arise, not from abstract rules, but from the operation of statistical learning mechanisms whose existence is uncontested. We do not claim that one of these possibilities is necessarily correct; our goal has simply been to point out that there are several alternatives to abstract, algebraic rules, and that the results do not implicate such rules because they provide no differential support for abstract rules relative to the other alternatives.

Conclusion

Generalization of knowledge from given examples to new cases is crucial for intelligent behavior; as Marr[14] pointed out, experience never repeats itself, and so our reactions to every experience depend to some degree on generalization. Marcus and his collaborators are right to emphasize the importance of generalization, and the experiments they have reported likely reflect the existence of impressive powers of generalization in infants. We have suggested, however, that some participants in the debate about the need for rules may have underestimated the potential of alternative forms of computation to address the problem of generalization by mistakenly assuming that statistical learning procedures, including neural networks, are doomed to compute statistics only over 'given variables'[4]. In fact neural networks make extensive use of internal representations, onto which the given variables (i.e. the raw input) are mapped. What sets some of the most interesting types of statistical learning procedures often used with neural networks apart from older (and for some, more familiar) statistical procedures is the fact that the network procedures can learn what internal representations ought to be assigned to the given variables. It seems likely to us that infants are born with predispositions to encode inputs in particular ways and with powerful statistical learning procedures like those currently used in network models that can help them refine their initial predispositions and discover new ones. As far as we can tell, there is no evidence to suggest that such procedures are insufficient to account for the sort of generalization seen in the Marcus *et al.* experiments.

Acknowledgements

Supported by NIH grants MH-47566 and MH-55628. We thank members of the CMU PDP research group for helpful comments and discussion.

References

1 Saffran, J.R., Newport, E.L. and Aslin, R.N. (1996) Statistical learning by 8-month-old infants *Science* 274, 1926–1928
2 Letters (1998) *Science* 276, 1177–1181
3 Marcus, G.F. *et al.* (1999) Rule learning by seven-month-old infants *Science* 283, 77–80
4 Putnam, H. (1995) Against the new associationism, in *Speaking Minds: Interviews with Twenty Eminent Cognitive Scientists* (Baumgartner, P. and Payr, S., eds), pp. 177–188, Princeton University Press
5 Pinker, S. and Prince, A. (1988) On language and connectionism: analysis of a parallel distributed processing model of language acquisition *Cognition* 28, 73–193
6 Kohonen, T. (1984) *Self-Organization and Associative Memory*, Springer-Verlag
7 Linsker, R. (1986) From basic network principles to neural architecture: I. Emergence of spatial-opponent cells *Proc. Natl. Acad. Sci. U. S. A.* 83, 7508–7512
8 Linsker, R. (1986) From basic network principles to neural architecture: II. Emergence of orientation-selective cells *Proc. Natl. Acad. Sci. U. S. A.* 83, 8390–8394
9 Linsker, R. (1986) From basic network principles to neural architecture: III. Emergence of orientation columns *Proc. Natl. Acad. Sci. U. S. A.* 83, 8779–8783
10 Miller, K.D., Keller, J.B. and Stryker, M.P. (1989) Ocular dominance column development: analysis and simulation *Science* 245, 605–615
11 Seidenberg, M.S. and Elman, J. Language learning: rules or statistics? *Science* (in press)
12 Dienes, Z.D., Altmann, G.T.M. and Gao, S-J. (1995) Mapping across domains without feedback: a neural-network model of transfer of implicit knowledge, in *Neural Computation and Psychology* (Smith, L.S. and Handcock, P.J.B., eds), pp. 19–33, Springer-Verlag
13 Dienes, Z.D., Altmann, G.T.M. and Gao, S-J. Mapping across domains without feedback: a computational model *Cognit. Sci.* (in press)
14 Marr, D. (1969) A theory of cerebellar cortex *J. Physiol.* 202, 437–470

Connectionism: with or without rules?

Response to J.L. McClelland and D.C. Plaut (1999)

Gary F. Marcus

G.F. Marcus is at the Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA.

tel: +1 212 998 3551
fax: +1 212 995 4292
e-mail: gary.marcus@nyu.edu

http://www.psych.nyu.edu/~gary

It is not altogether surprising that McClelland and Plaut, researchers with longstanding interests in providing alternatives to rules, find our recent experiments unconvincing [McClelland, J.L and Plaut, D.C. (1999) Does generalization in infant learning implicate abstract algebra-like rules? *Trends Cognit. Sci.* 3, 166–168][1]. But advocates of their cognition-without-rules view might want to look elsewhere to bolster their case, as none of McClelland and Plaut's objections turns out to be plausible.

Before addressing their objections, let me outline what I see as three important points of agreement. First, we all seem to be interested in the study of how cognition could be realized in a neural substrate. Second, we all believe that the study of neural networks can be helpful in this regard.

Third, we agree that a basic property of the class of models that McClelland and Plaut advocate is that they depend on the overlap of features. As they put it:

> generalization in neural networks depends on overlap of representations – that is, the patterns of activity used in the network – to represent items experienced during training and test. For prior learning to generalize to a new stimulus, the representation of the new stimulus must overlap with – that is, activate some units in common with – the representation of the stimuli on which learning is based. This is because learning occurs by the adjustment of connection weights between specific units in a network, and so a new input must activate some of the same units whose weights were influenced by prior experience to benefit from that experience. (Ref. 1, p. 166.)

As it turns out, I made almost exactly this point in a recent article[2]. Where we seem to disagree is with the implications of this fact about overlap. The problem, as I see it, is that this inability to generalize to non-overlapping items renders a certain class of network models inappropriate for many cognitive tasks, because in many cognitive tasks we are required to generalize to new items that do not

1364-6613/99/$ – see front matter PII: S1364-6613(99)01321-2