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 overlap with the items that we have seen before.

The experiments in our *Science* article[3], discussed by McClelland and Plaut, were designed to address precisely this point, testing whether infants could generalize an abstract relation, such as the one found in 'ABA' grammar consisting of sentences such as '*li-ti-li*' and '*ga-ti-ga*', to novel items that didn't overlap with previous items. Infants were able to do this, discriminating consistent test sentences like '*wo-fe-wo*' from inconsistent test sentences like '*wo-fe-fe*'. If words are represented as independent items, as they are in, say, Elman's work on the popular simple recurrent network[4], the test items do not overlap with the habituation items, thus the standard version of the simple recurrent network does not succeed in making the discrimination. This is exactly as would be expected, given McClelland and Plaut's discussion about the importance of overlap. (This is an entirely replicable result, using a wide range of network parameters. Readers who wish to verify this for themselves can look at the sample files we provide at http://www.psych.nyu.edu/~gary/es.html.)

Furthermore, our experiments were designed in such a way that even if the input were encoded as sets of phonetic features a standard simple recurrent network would still be unable to discriminate the consistent and inconsistent test stimuli. Again this follows because the relevant features that would discriminate consistent from inconsistent test items did not overlap with what the model would have learned about in the habituation.

Of course, what counts as 'overlapping' depends on how inputs are encoded. The words *cat* and *dog* would presumably overlap if they were represented as sets of semantic features, but not if they were represented in terms of their orthographic (i.e. spelling) features.

Auditory contours: an alternative account?

It is in this context that McClelland and Plaut point out, quite rightly, that infants could encode our stimuli in other ways, for example, in terms of sound contours like '+loud, –loud, +loud'. If it were the case that the test items overlapped in terms of those sound contours, infants could (in principle) succeed using a standard simple recurrent network that used those features as inputs (although one could quibble about whether the model could do so fast enough, etc.). Of course, this alternative would only work if the relevant sound contours are available in our data, and if the child encodes the linguistic-like stimuli using those contour features. But McClelland and Plaut did not actually test our materials to see if these contours were present. As it turns out, the relevant differences (between *bo* and *po* and between *ko* and *ga*) appear to be less than one decibel, and hence unlikely to be discernible.

In any event, even if the differences were somewhat larger, we doubt that a child would use them. Words vary in their loudness and pitch all the time, but for the most part we seem to filter out that variability: no language learner should treat the word *cat* differently depending on whether it is spoken at 62 decibels or 63. Likewise, excepting tonal languages, we would not expect a language learner to treat words differently depending on slight variations in pitch.

Still, on the maxim that it is more convincing to counter argument with data than with further argument, we have collected preliminary data from six infants in a follow-up experiment. In this experiment, infants were again trained on either AAB or ABB sentences, but we changed the second word of each test sentence such that it was noticeably different in loudness (by about 2.5 db) and pitch from the first and third words. If infants were relying on the sound contours, the effects in our original experiments would disappear in this version. Instead, the results appear to be unchanged: five of our six subjects looked longer at the syntactically inconsistent test items than at the syntactically consistent items, in line with what we found in our earlier work. This is exactly as it should be, for what matters in language (tonal languages and stress aside) is not how loudly somebody says something, nor the fundamental frequency of their voice, but rather what words they are saying and what the relationship is between those words.

Statistics

Elsewhere McClelland and Plaut appear to broaden the notion of statistics from things like transitional probabilities between particular elements to any kind of relation between any kind of information, concrete or abstract. The trouble is that this broader notion of statistics trivializes the very term, rendering it broad enough to encompass any lawful relationship, including for example, the very rules that McClelland and Plaut argue against. For example, by the definition of statistics that McClelland and Plaut implicitly adopt, if a language produced sentences only of the type noun-phrase followed by verb-phrase, one could describe the language in terms of a phrase-structure rule [Sentence→Noun-Phrase, Verb-Phrase], but also in terms of a statistical pattern in which verb phrases follow noun phrases one hundred percent of the time. We did not mean to deny that children could make use of statistics in this broader, probably unfalsifiable sense; our intent was only to argue against the narrower definition of statistics. Our criticism of models that rely purely on transitional probabilities between words still holds, and we do not see a proposal for a kind of statistical reasoning that would succeed in our task without (perhaps covertly) encompassing rules.

It is also, of course, fine to have some external device compute whether any two items are the same, and then compute the statistical likelihood that that external same–different system will say 'yes'. But if that external system itself implements a rule (e.g. a line of computer code that says, for all syllables x, y, if x equals y execute condition A, otherwise execute condition B), we are still left with a system that incorporates a rule. Relocating a rule is *not* tantamount to eliminating it.

Learning the learning mechanism

Another idea worth considering is McClelland and Plaut's suggestion that the learning mechanism itself could be learned: '...powerful mechanisms might simply be ones that help statistical learning procedures generalize in powerful ways. Furthermore, these mechanisms might themselves be learned.' While we agree that it is a logical possibility that some learning mechanisms might themselves be learned, we note that (1) no such proposal has actually been made, and (2) there must be a solution to the bootstrapping problem; which is to say that on pain of infinite regress, learning can only take place in a system in which at least some learning mechanism is innate.

Models

In the remainder of their critique, McClelland and Plaut focus on connectionist models, attacking a claim that we never made. We never intended to deny that one could build a neural network that could capture our data. Rather we aimed 'to try to characterize what properties the right sort of neural network architecture must have'. Here and elsewhere the difference between different kinds of neural network models has been obscured, as though all networks were alike, and as if the success of a given network model automatically counted against the rule hypothesis. But networks are not in fact all alike – some implement rules (overtly or covertly), some do not. Our work aimed to provide a mechanism for choosing between different types of models; as we shall see, the models that work the best are those that implement, rather than eliminate, rules.

Seidenberg and Elman

For example, consider the recent model of Seidenberg and Elman mentioned by McClelland and Plaut. To some extent this model can capture our data (McClelland and Plaut concede that the model is not perfect, writing that 'one may quibble with the particulars of the reported simulation'). But this model turns out to depend on a behind-the-curtain 'teacher' that itself incorporates a rule. (In this case, the rule – that is, operation over variables that can be applied to any instance – was probably implemented as a line of computer code such as 'if a = b then output 1 else output 0'.)

In this respect, the Seidenberg and Elman model is a significant departure from Elman's earlier work, abandoning his commitment to 'prediction tasks'. As Elman himself noted[5] the virtue of the prediction task is that 'One [issue] which arises with supervised learning algorithms such as backpropagation of error is the question of where the teaching information comes from. In many cases, there are plausible rationales which justify the teacher. But the teacher also reflects important theoretical biases which one might sometimes like to avoid' whereas 'the prediction task... represents information which is directly observable from the environment.' In contrast, the Seidenberg and Elman model depends on an internal teacher that must do some computation on the information provided in the environment. There is nothing wrong with such a *deus ex machina* but it is crucial to realize that the *deus ex machina*, which itself depends on a rule, must be taken as part of the system as a whole, and to realize that without that rule, the whole system breaks down. As I have written elsewhere in a reply to Seidenberg and Elman, they 'have not eliminated the rule, they have simply hidden it'[6].

Dienes and Altmann

Finally, let us turn to the model by Dienes, Altmann and Gao[7] that McClelland and Plaut advocate. Unlike the Seidenberg and Elman model, Dienes and Altmann's architecture does not include an external teacher that builds in a sameness-detecting rule. Instead, the Dienes *et al.* model instantiates a different hypothesis about transfer, one in which words in a second vocabulary are mapped onto words in a first vocabulary. McClelland and Plaut speculate that this model might be able to capture our data.

We have not yet had time to analyze fully whether this model can in fact capture our data, but in our preliminary experiments with the model we have found the following: if the model is trained on our habituation sentences (e.g. '*la-ta-la*' and the like) and then tested on many consistent test sentences with a new vocabulary, such as '*wo-fe-wo*', and then subsequently tested on '*fe-wo-wo*' versus '*fe-wo-fe*', the model is more 'surprised' to hear '*fe-wo-fe*' (consistent) than '*fe-wo-wo*' (inconsistent). We suspect that this is a consequence of mapping the second vocabulary onto the first, but we doubt that children would do the same, and plan to test this prediction of the model.

Discussion

None of this is to say that you cannot build a connectionist model that can capture our results. The property of generalizing only to overlapping items is not intrinsic to neural networks; it is possible to build neural networks that do not have that property. As we noted in our article, Holyoak and Hummel[8,9] had already done so, building a model that captures a task that is equivalent to ours; Shastri and Chang (pers. commun. and Refs 10,11) have now done so as well. But these authors embrace rules rather than scorning them, implementing explicit variables and abstract relationships between variables. From the perspective of comparing a broad range of possible models, it is unfortunate that McClelland and Plaut do not even address the kinds of model that Holyoak and Hummel and Shastri and Chang advocate.

McClelland and Plaut worry that they 'don't really see how experiments' like ours can tell us 'whether [infants] use rules' – without suggesting any alternative. We find such a view to be unduly pessimistic, casting questions about models as unanswerable. While we acknowledge the fact that it is impossible to test the broad framework of connectionism – which encompasses both systems that use rules and those that do not – it is possible to use empirical data to choose between classes of models, and we believe that our experiments do so. Our data are not readily captured by models that do not incorporate rules (such as the original version of the Simple Recurrent Network) but work by Holyoak and Hummel, and Shastri and Chang, has shown that our results can be captured in a variety of models, including connectionist models that do incorporate rules.

Acknowledgement

We thank Zsofia Zvolensky for a suggestion that led to the '*fe-wo-fe*' test of the Dienes *et al.* model.

References

1 McClelland, J.L. and Plaut, D.C. (1999) Does generalization in infant learning implicate abstract algebra-like rules? *Trends Cognit. Sci.* 3, 166–168

2 Marcus, G.F. (1998) Rethinking eliminative connectionism *Cognit. Psychol.* 37, 243–282

3 Marcus, G.F. *et al*. (1999). Rule learning in 7-month-old infants *Science* 283, 77–80

4 Elman, J.L. (1990) Finding structure in time *Cognit. Sci.* 14, 179–211

5 Elman, J.L. (1995) Language as a dynamical system, in *Mind as Motion: Explorations in the Dynamics of Cognition* (Port, R.F. and van Gelder, T., eds), pp. 195–223, MIT Press

6 Marcus, G.F. Reply *Science* (in press)

7 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 Hancock, P.J.B., eds), pp.19–33, Springer-Verlag

8 Hummel, J.E. and Holyoak, K.J. (1997) Distributed representations of structure: a theory of analogical access and mapping *Psychol. Rev.* 104, 427–466

9 Holyoak, K.J. and Hummel, J.E. The proper treatment of symbols in a connectionist architecture, in *Cognitive dynamics: Conceptual change in humans and machines* (Deitrich, E. and Markman, A. eds), Erlbaum (in press)

10 Shastri, L. and Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behav. Brain Sci.* 16, 417–494

11 Shastri, L. (1997) *Exploiting Temporal Binding to Learn Relational Rules Within a Connectionist Network* (TR-97-003), International Computer Science Institute, University of California, Berkeley