Memory & Cognition
1980, Vol. 8 (4), 378-382

Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model

DOUGLAS L. HINTZMAN and GENEVIEVE LUDLAM
University of Oregon, Eugene, Oregon 97403

A common finding in studies of classification learning is that ability to classify the prototype of a category declines much less over a retention interval than does the ability to classify the previously seen exemplars themselves. We demonstrate here that this finding does not necessarily indicate the existence, in memory, of a representation of the prototype. MINERVA, a computer-simulation model that encodes memory traces only of presented exemplars, was tested on an appropriate task. Differential forgetting of prototypes and old instances was shown by a version of the model that assumed that (1) classification is based on the exemplar trace most similar to the test stimulus and (2) individual properties are lost from the traces over time in an all-or-none fashion. It is suggested that, in general, the key to the prediction of differential forgetting may be the concomitance of forgetting and generalization.

One way for a device to learn to classify stimuli is to develop a prototype of each category representing the "central tendency" of instances that have been encountered and to assign each new stimulus to the category whose prototype it most closely matches. It has been claimed that, at least in certain tasks, humans learn classification in this way. The most compelling evidence for this conclusion has come from studies comparing changes over a retention interval in classification performance on exemplars and on the prototypes from which the exemplars were derived. Specifically, several studies have shown that the ability to classify old exemplars declines substantially over time, whereas the ability to classify prototypes, not seen during original learning, declines little and in some cases may even improve (Homa, Cross, Cornell, Goldman, & Schwartz, 1973; Posner & Keele, 1970; Strange, Kenney, Kessel, & Jenkins, 1970). This differential forgetting suggests that the prototype has a representation of its own. If the prototype were classified according to its similarity to traces of the individual instances, the argument goes, then classification of the prototype would decline as rapidly as classification of the instances themselves. Acceptance of this argument, in addition to confirming a prototype view of classification learning, might be interpreted as supporting the distinction between episodic and generic memories. It would identify the two types of memory with different forgetting rates (Hintzman, 1978, pp. 374-376).

The argument, however, is not necessarily correct. We present here the results of a computer simulation of classification learning, using a model in which only old instances, and not prototypes, are stored. The model is capable of predicting differential forgetting rates for prototypes and old exemplars, providing that certain assumptions are met. We shall give special consideration to the conditions that are necessary to produce this result.

This material is based upon work supported by the National Science Foundation under Grant BNS-7824987. Requests for reprints should be sent Douglas L. Hintzman, Psychology Department, University of Oregon, Eugene, Oregon 97403.

THE MODEL

The simulation model, dubbed MINERVA, was developed as the first stage of a test of adequacy of a multiple-trace account of the effects of repetition on memory. The theory is similar to that of Semon (1923; see also Schacter, Eich, & Tulving, 1978) in assuming that each individual experience gives rise to its own memory trace and that performance that may suggest that a separate, generic representation has been abstracted out of the individual experiences is actually based on the retrieval of these episodic traces, either singly or in concert. As applied to the classification learning task, the theory is most similar to those of Brooks (1978) and Medin and Schaffer (1978), in that it denies that exposure to exemplars gives rise to a generic prototype representation, and it assumes that only the traces of individual exemplars are stored. Classification is accomplished through a version of what Reed (1972) calls the "proximity algorithm"; that is, the one or more traces most closely matching the new stimulus determine how the stimulus will be classified. In the following description of the model, we have attempted to convey enough details so that the reader will understand how the simulation was done. The discussion will be concerned with those characteristics of the model necessary to produce differential forgetting of prototypes and old exemplars, which is our focus of interest here.

0090-502X/80/040378-05$00.75/0

All experiences are assumed to be configurations of primitive elements or properties. The set of primitive elements includes not only features (e.g., blue, triangle), but also relations (e.g., larger than, inside). Both new stimuli and memory traces are represented as configurations of these properties. In the program itself, a list of allowable properties relevant to the task at hand is provided by the programmer at the beginning of each simulation run.

When a stimulus is input, each of its properties is assigned a strength (S). In the following example and in the simulations reported here, S = 1.0 has arbitrarily been assigned to features and S = .5 to relations. Note that in the complete description of the stimulus, a relation and its inverse (e.g., larger than, smaller than) are both represented. A typical stimulus description is as follows. Stimulus 2: category = A; number of objects = 2 (e and f); properties of Object e–size = larger than f (S = .5), color = yellow (S = 1), shape = square (S = 1), position = above f (S = .5); properties of Object f–size = smaller than e (S = .5), color = red (S = 1), shape = triangle (S = 1), position = below e (S = .5).

The initial encoding of a memory trace is simply a copy of the stimulus description. Each primitive property employed in a trace description is linked from its representation in the list of allowable properties to the appropriate "object" in the memory trace (e.g., in the above example, Object e or f).

A new input is matched for similarity against all traces in memory. (In the theory, this is assumed to be done in parallel; in the simulation, of course, it is done serially.) Call the input I and a given trace T. Then S_I and S_T are the strengths of a given property in the input and the trace, respectively. The degree of match between the input and the trace is computed by a modified version of Tversky's (1977) formula:

$$M(I,T) = \sum^{(I\cap T)} S_I \cdot S_T - \sum^{(I-T)} S_I^2 - \sum^{(T-I)} S_T^2.$$

Here, I∩T represents properties shared by I and T, I–T represents properties in I but not T, and T–I represents properties in T but not I. The match is determined "object by object." Thus, in order for a perfect match to occur, it is not enough that all the same features and relations exist in I and in T, they must also have the same configuration (e.g., the yellow square must be above the red triangle, and not vice versa). That is to say, if the relation is the same but the configuration is not, this relational property will fall in sets I–T and T–I, and not in I∩T.

It is assumed that a retrieval cue will retrieve only traces with which it has considerable overlap and, therefore, that poorly matching Ts will not influence the way I is classified. There is no a priori basis for deciding how good a match must be in order to be involved in classification; however, we have arbitrarily drawn the line at zero in the above formula. That is, all of the classification procedures to be described operate only on positive values of M(I,T).

We have compared four ways of predicting the accuracy of classification of an input, I. Assume that the correct category is A and the incorrect category is B. The four predictors of classification are: (1) A average, the mean of the positive $M(I,T_A)$ values, (2) A–B average, the difference between the mean of the positive $M(I,T_A)$ values and the mean of the positive $M(I,T_B)$ values, (3) A best, the maximum $M(I,T_A)$ value, and (4) A–B best, the difference between the maximum $M(I,T_A)$ and the maximum $M(I,T_B)$. Procedures 1 and 2 make use of central tendencies, and in this respect they are like prototype models. However, the "prototype" here does not have a unitary representation; it is computed at the time of retrieval. More important, it is the central tendency of positive M(I,T) values only; Ts dissimilar to I do not influence how I is classified. Thus, Procedure 2 uses what Reed (1972) termed the proximity algorithm, but in this case with the number of contributing traces variable, rather than fixed. Procedures 3 and 4 are best-match procedures, similar to those advocated by Brooks (1978) and by Medin and Schaffer (1978). Procedure 4 is, essentially, the proximity algorithm, with the number of contributing traces per category equal to one. In all four procedures, however, only positive M(I,T) values contribute. If there are no such values in a given category, both the mean and maximum M(I,T) values default to zero.

We have also compared two ways of producing forgetting. In one, each S value in a trace is decremented by a certain proportion of its present value on each forgetting cycle [i.e., $S_{t+1} = (1-\theta)S_t$, where $0 \leqslant \theta \leqslant 1$]. In the other, on each cycle, a property is deleted from the trace in an all-or-none fashion with a given probability, F. The first of these, decremental forgetting, does not affect the relative numbers of properties falling in the sets I∩T, I–T, and T–I. The second, however, does. When forgotten, a property originally in I∩T shifts to I–T; one originally in T–I in effect disappears.

PRELIMINARY TESTS

Medin and Schaffer (1978) performed several experiments designed to differentiate an exemplar-based theory of classification learning from theories that assume that classification is based upon representations of prototypes. In each experiment, stimuli varied on four binary dimensions. Training stimuli were divided into two categories, A and B, in such a way that the prototype and exemplar-based theories made clearly different predictions concerning both the relative difficulties of acquisition of the training stimuli and the classification of new transfer stimuli that the subjects had not seen before. In all the Medin and Schaffer experiments, their exemplar-based theory fit the data very well.

Although the present model differs from that of Medin and Schaffer (1978) in a number of ways that we will not go into here, we expected it to behave similarly, particularly when classifying on the basis of best match. Accordingly, simulations were run of Medin and Schaffer's (1978) Experiment 1. Data were generated for 32 simulated subjects, with one cycle of all-or-none forgetting (F = .25) prior to testing, to simulate imperfect acquisition. Correlations between each of our classification procedures and the mean confidence ratings obtained by Medin and Schaffer, for training and transfer stimuli combined, were: A average, r = .73; A–B average, r = .84; A best, r = .82; A–B best, r = .88. A similar simulation of Medin and Schaffer's Experiment 2 (using F = .35) yielded: A average, r = .70; A–B average, r = .82; A best, r = .78; A–B best, r = .85. Thus, the ordering of the three classification procedures was the same for both experiments, with A–B best giving the best fit in both cases. These fits are not as good as those of the Medin and Schaffer model, which has several free parameters, and the fits could no doubt be improved by allowing S values for different attributes to vary, as a means of taking dimensional salience in account. Such an exercise, however, would be costly and trivial. The important point to note is that, using the A–B best classification rule, the model makes essentially the same predictions as the Medin and Schaffer (1978) theory. This provides added evidence for their conclusion that the classification performance of their subjects was based on individual stored instances or exemplars, rather than on representations of prototypes.

EFFECTS OF FORGETTING ON THE CLASSIFICATION OF EXEMPLARS AND PROTOTYPES

To simulate a typical experiment on the forgetting of prototypes, two prototype stimuli were selected and seven exemplars were generated from each one. The prototypes, each having seven independent binary dimensions, were A, a large blue square above a small red triangle, and B, a large red circle above a small yellow triangle. Seven exemplars of each category were then generated by changing the value of one dimension at a time. Relational properties, such as above-below, which are necessarily redundant, were treated as unitary. (As in the example given earlier, the two aspects of each such relational property were given initial strengths of S = .5, rather than S = 1.)

For each simulated subject, the classifications of 12 stimuli (six exemplars from each category) were stored in memory. These will be referred to as the old exemplars. The two remaining exemplars (the new exemplars), one per category, were used in testing only, as were the two prototypes. The exemplars held out as new were rotated within each category, producing seven different sets of stimuli. Seven simulations were run, one using each stimulus set. Forgetting was all or none, with F = .25. Each simulation included one test with no forgetting, one test after one iteration of the forgetting procedure, and one test after three. All 16 stimuli were tested each time, and all four classification measures were computed.

Mean values for old exemplars, new exemplars, and prototypes are shown for each measure in Figure 1. When classification was based on average matches (A average and A–B average) prototypes were classified better than old exemplars, and old exemplars were classified better than new exemplars, and this ordering did not change as a function of number of forgetting cycles (time). The reasons for the ordering should be obvious: Prototypes are similar to all stored exemplars and, therefore, match well on the average. The trace of an old exemplar contributes to its average match, but there is no trace to provide this advantage to a new exemplar, and thus classification of new exemplars is poorest of all.

The two best-match measures (A best and A–B best) show a different pattern, one very much like that produced by human subjects (Homa et al., 1973; Posner & Keele, 1970; Strange et al., 1970): Old exemplars may be classified best on an immediate test, but they lose their advantage over time. Performance on prototypes changes little as a function of time and may exceed that on old exemplars after a long retention interval. Performance on new exemplars lies below and parallels that on prototypes.

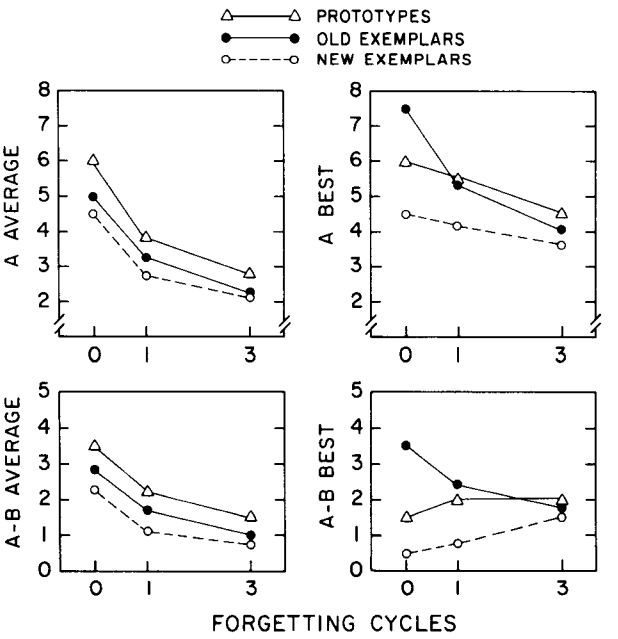

Figure 1. Effects of zero, one, and three iterations of the all-or-none forgetting procedure on classification of old exemplars, new exemplars, and prototypes according to four predictors of performance.

Why did this interaction emerge from the simulation using best-match measures? Classification of an old exemplar declines rapidly because, in the beginning, it matches one trace (its own) the best, but the more times the forgetting procedure is invoked, the poorer that match becomes. The prototype matches six stored exemplars fairly well, but not perfectly. As properties are deleted from these traces in a random fashion, the traces match the prototype, on the average, less and less well. But if classification is based on the best match, rather than the average match, then it is likely that at least one of the six traces will still match the prototype fairly closely. Indeed, if a trace loses only the one property that differentiates it from the prototype, its similarity to the prototype, as measured by M(I,T), will actually increase. The forgetting curve for the prototype crosses that for old exemplars, then, because the prototype has a statistical advantage. As information is lost, it is better to be close to all six original exemplars than to be identical to only one. The explanation for the new exemplar curve is the same as that for the prototype curve, except that the new exemplar is more unique than the prototype; it is no more similar to the old exemplars than the old exemplars are to each other. The new exemplar and old exemplar curves, therefore, should approach the same value, but never cross.

We have not presented here data resulting from the decremental forgetting procedure, which simply decreases the strengths assigned to properties, because preliminary investigations showed no interesting interactions. Differential forgetting of prototypes and old exemplars, given the trace structure and matching function we have employed, appears to require that forgetting of properties be all or none. That is, merely decreasing the strengths of properties does not make them jump from I∩T to I–T, or from T–I to oblivion, and such jumps seem to be necessary to produce the obtained effects. Presumably, however, a strength-plus-threshold model could mimic the behavior of the model that was successful here.

DISCUSSION

The version of MINERVA that was most successful on these classification tasks is similar to several other theories in the literature. Like the theories of Brooks (1978) and Medin and Schaffer (1978), it assumes that only individual exemplars are represented in memory. MINERVA must be seen as a member of the class of models in which relationships among features, not simply the features themselves, are involved in classification (cf. Medin & Schaffer, 1978), because classification is based on exemplars, and each exemplar is a conjunction of features.

The successful A–B best rule, as was noted earlier, is essentially the same as Reed's (1972) proximity algorithm, in the special case in which classification is determined by the single exemplar trace that is most like the test stimulus. It is important to note that this rule does not mean that the test stimulus will be automatically assigned to the category to which the most similar exemplar belongs, because the loss of information in memory can make the trace of that exemplar less like the test stimulus than is the trace of some other exemplar. The assumption of all-or-none forgetting of features borrows from Bower's (1967) multicomponent theory of the memory trace, and this suggests that some version of that theory would predict the behavior shown in the present simulations.

What we have demonstrated most clearly here is that differential forgetting of prototypes and old exemplars, as shown by Homa et al. (1973), Posner and Keele (1970), and Strange et al. (1970), cannot be taken as conclusive evidence for the abstraction of a prototype during classification learning, for there is at least one set of assumptions that predicts this result purely from the forgetting of old exemplars, with no ad hoc mechanisms required. Our findings thus provide added support for the notion that classification learning may be based solely on memory for past instances, without involving the representation of a prototype or abstract idea.

In part, this explanation of differential forgetting might have been anticipated on the basis of past empirical findings alone. In a recognition memory task, Bahrick, Clark, and Bahrick (1967) showed that over a retention interval, even as the frequency of hits is falling, the frequency of false alarms to stimuli very similar to the target can increase; that is, generalization increases over time. A distractor item very similar to several targets (i.e., a "prototype") would, of course, benefit from several overlapping generalization gradients and presumably could, after sufficient time, seem more "familiar" than any of the original targets themselves. Such a prediction is easily extended from recognition memory to the classification task.

These considerations suggest that the key to predicting differential forgetting of prototypes and old exemplars is not all-or-none forgetting or the proximity algorithm per se, but the concomitance of forgetting and generalization. Encoding, forgetting, retrieval, and matching assumptions interact, and so other combinations of assumptions might be expected to mimic the behavior of MINERVA. Medin and Schaffer (1978) have proposed that, in contrast to the present view, generalization might be the primary and forgetting the secondary process (see also Gibson, 1940). While this hypothesis might be rejected on other grounds (cf. Underwood, 1961), it seems possible that some version of the Medin and Schaffer theory would predict the differential forgetting results.

The present results do not, of course, show that prototype theories are wrong. But they do show that the acceptance of such theories should not rest on the differential forgetting result alone. Other, converging

evidence is needed to support prototype theory. One approach to providing such evidence has been explored by Robbins, Barresi, Compton, Furst, Russo, and Smith (1978), who showed that in learning the reversal of exemplar-response pairs, subjects reversed immediately after original learning made many more errors on changed than on unchanged pairings, while those reversed after a long delay found changed and unchanged pairings about equally difficult to learn. This is what would be expected if the basis of performance were changing, over time, from traces of individual instances to a representation of a prototype; Robbins et al. interpreted their data in this way. Unfortunately, the same finding would be expected if, over time, subjects became less able to distinguish among instances, and thus less able to tell which pairings had changed and which had remained the same. Thus, unequivocal support for a prototype theory as opposed to an exemplar-based theory of classification learning still appears to be lacking.

REFERENCES

Bahrick, H. P., Clark, S., & Bahrick, P. Generalization gradients as indicants of learning and retention of a recognition task. *Journal of Experimental Psychology,* 1967, **75**, 464-471.

Bower, G. H. A multicomponent theory of the memory trace. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 10). New York: Academic Press, 1967.

Brooks, L. Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization.* Hillsdale, N.J: Erlbaum, 1978.

Gibson, E. J. A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review,* 1940, **47**, 196-229.

Hintzman, D. L. *The psychology of learning and memory.* San Francisco: Freeman, 1978.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Schwartz, S. Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology,* 1973, **101**, 116-122.

Medin, D. L., & Schaffer, M. M. Context theory of classification learning. *Psychological Review,* 1978, **85**, 207-238.

Posner, M. I., & Keele, S. W. Retention of abstract ideas. *Journal of Experimental Psychology,* 1970, **83**, 304-308.

Reed, S. K. Pattern recognition and categorization. *Cognitive Psychology,* 1972, **3**, 382-407.

Robbins, D., Barresi, J., Compton, P., Furst, A., Russo, M., & Smith, M. A. The genesis and use of exemplar vs. prototype knowledge in abstract category learning. *Memory & Cognition,* 1978, **6**, 473-480.

Schacter, D. L., Eich, J. E., & Tulving, E. Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior,* 1978, **17**, 721-743.

Semon, R. [Mnemic psychology] (B. Duffy, trans.). London: George Allan & Unwin, 1923.

Strange, W., Kenney, T., Kessel, F., & Jenkins, J. Abstraction over time of prototypes from distortions of random dot patterns. *Journal of Experimental Psychology,* 1970, **83**, 508-510.

Tversky, A. Features of similarity. *Psychological Review,* 1977, **84**, 327-352.

Underwood, B. J. An evaluation of the Gibson theory of verbal learning. In C. N. Cofer (Ed.), *Verbal learning and verbal behavior.* New York: McGraw-Hill, 1961.

(Received for publication February 8, 1980;
revision accepted April 23, 1980.)