Similarity and Rules United: Similarity- and Rule-Based Processing in a Single Neural Network

Tom Verguts, Wim Fias

Department of Experimental Psychology, Ghent University

Received 26 November 2007; received in revised form 17 July 2008; accepted 23 July 2008

Abstract

A central controversy in cognitive science concerns the roles of rules versus similarity. To gain some leverage on this problem, we propose that rule- versus similarity-based processes can be characterized as extremes in a multidimensional space that is composed of at least two dimensions: the number of features (Pothos, 2005) and the physical presence of features. The transition of similarity-to rule-based processing is conceptualized as a transition in this space. To illustrate this, we show how a neural network model uses input features (and in this sense produces similarity-based responses) when it has a low learning rate or in the early phases of training, but it switches to using self-generated, more abstract features (and in this sense produces rule-based responses) when it has a higher learning rate or is in the later phases of training. Relations with categorization and the psychology of learning are pointed out.

Keywords: Similarity; Rules; Learning processes; Neural networks

1. Introduction

Any 2-year-old child confronted with a dolphin rapidly classifies it as a big fish. Many years and biology classes later, however, the same child will call the same animal a mammal, and not a fish. This example illustrates what experimental psychologists have documented with careful experimentation: some types of reasoning are similarity-based, and others are rule-based (e.g., Allen & Brooks, 1991; Hampton, 1995; Rips, 1989; Shanks & Darby, 1998; Sloman, 1996; Smith, Patalano, & Jonides, 1998). The distinction is undisputed at the behavioral level, but it is much more controversial whether the behavioral distinction necessitates fundamental architectural differences underneath the two types of
reasoning (Marcus, Vijayan, Rao, & Vishton, 1999; Pothos, 2005; Seidenberg & Elman, 1999; Sloman, 1996). While there is general consensus about the existence of a similarity-based processing system in both humans and animals, the existence of an additional and qualitatively different rule-based system is controversial (in humans and possibly in animals; Beckers, Miller, De Houwer, & Urushihara, 2006; Penn & Povinelli, 2007). Rules-versus-similarity controversies have been discussed in many domains including, but not restricted to, categorization (Nosofsky, 1984; Rips, 1989), reasoning (Landy & Goldstone, 2007; Sloman, 1996), the psychology of learning (Pearce & Bouton, 2001; Shanks, 2007), and language (McClelland & Patterson, 2002; Pinker & Prince, 1988).

The aim of the work presented in this paper is to provide a proof of principle that a single cognitive architecture with a single learning mechanism can develop both similarity-based and rule-based behavior. Although similar claims have been reported earlier in the neural network literature (e.g., Rumelhart & McClelland, 1987), the scope of the present paper is considerably different from these earlier models, as will become clear below. We start from the conceptualization that similarity-based and rule-based behavior are prototype concepts in a multidimensional space consisting of dimensions such as number of features used in categorization, abstractness of features, and others. Depending on different characteristics (e.g., learning speed, number of trials, type of task) a model can take many different positions in this space, and move during training from similarity- to rule-based processing or vice versa. The view that a single system suffices does not of course imply that there is a single processing route for a given task; it only denies that qualitatively different mechanisms underlie different processing routes.

From the diverse instances where similarity- and rule-based behavior have been reported, we selected a task paradigm in which the behavioral dissociation between the two types of reasoning was as clear-cut as possible. For this purpose we chose the paper of Shanks and Darby (1998) who found that different participants responded in one of two completely different ways to the same stimuli; one way was interpreted as similarity based and the other as rule based (we only discuss their Experiment 2, as it is a more fully balanced version of their Experiment 1). Shanks and Darby asked their participants to imagine they were medical doctors investigating hypothetical patients eating different foods (labeled A, B, ...). Each patient developed an allergic reaction (+) or not (−) after eating a particular food or food combination (e.g., A+, patient had eaten A and obtained an allergy; or AB−, patient had eaten A and B and did not develop an allergy). The experiment was organized in a training phase followed by a test phase. In the training phase, some single food items (single-element stimuli) were associated with a particular allergy (A+, B+, E+, F+), whereas combinations of these food items (compound stimuli) were not associated with the allergy (AB−, EF−). Conversely, other single food items were not associated with an allergic reaction (C−, D−, G−, H−), whereas combinations of these food items led to an allergic reaction (CD+, GH+). Critically, to test generalization, Shanks and Darby also presented single-element stimuli I+, J+, M−, and N− without the compound stimuli IJ and MN in the training phase, and similarly, compound stimuli KL− and OP+ without the corresponding single-element stimuli K, L, O, and P.
From the stimuli presented in the training phase, a rule-based reasoner might infer the regularity: “A compound and its elements predict opposite outcomes.” Hence, a rule-based reasoner would infer that the compound IJ does not evoke the allergy given that the single food items did evoke an allergic reaction (IJ− because of I+ and J+ in the training phase); conversely, a rule-based reasoner would conclude that the compound MN does evoke the allergy (MN+) given that the single food items did not (M− and N−). Also, a rule-based reasoner would infer that K and L separately are associated with the allergy (K+ and L+ because of KL− in the training phase), and O and P are not (O− and P−). On the other hand, a similarity-based reasoner would come to the opposite conclusions. In particular, due to the similarity of IJ to the single-element stimuli I and J, a similarity-based reasoner would infer IJ+; and a similar argument applies to the other combinations (MN−, K−, L−, O+, and P+).

Participants were divided into three groups depending on their accuracy in the training phase. This grouping was partly manipulated experimentally; some participants received lots of training (which increases the probability of high accuracy in the training phase), some received intermediate levels of training and other participants received less training. Accuracies in the final training block were (with a maximum of 18, for 18 training stimuli), 12.1 (standard deviation [SD] = 1.8), 16.5 (SD = 0.6), and 18 (SD = 0) in the Low, Medium, and High group, respectively. The test phase results are shown in Fig. 1. There are two main things to be noted in these data. First, there are two generalization strategies to the transfer compound stimuli (IJ, MN; Fig. 1A). One group of participants with high training performance responds to the compound stimuli in the test phase in a way that is opposite to the single-element associations acquired in the training phase (“allergy” to MN but not to IJ; High group, rule-based), whereas the other group does not take into account the reversal rule and tends to respond in the same way to the single-element stimuli as to the compounds (Low group, similarity-based). There is a suggestion of a similar trend for the situation where the participants

![Figure 1](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAgAAAAAQAQMAAABg0K1wAAAABGdBTUEAALGPC/xhBQAAAABlBMVEX///8AAAG1JREFUeNrs9bKQdCgD3/8AAAABJRU5ErkJggg==)

Fig. 1. Data of Shanks and Darby (1998), Experiment 2, test phase. (A) Transfer compound stimuli. (B) Transfer single stimuli. The notation K/L (O/P) denotes the average percent allergy predictions to the stimuli K and L (O and P).
were exposed to the compound stimuli in the training phase and have to respond to the single-element stimuli in the test phase (K/L, O/P; Fig. 1B), although the data pattern is much more ambiguous (the interaction is reported by Shanks and Darby to be marginally significant in Fig. 1B; \( p = .065 \)). Second, which strategy a participant uses depends on his/her training performance; persons with high training performance tend to use a rule-based reasoning strategy, persons with low training performance tend to use a similarity-based reasoning strategy, as can be seen from the interaction in Fig. 1A. We provide a conceptual interpretation of these data in the next section to pave the way for an integrative model.

2. Integrating similarity- and rule-based processing

Although the Shanks and Darby (1998) experiment seems to constitute compelling evidence for two qualitatively different modes of reasoning, a unified view of the two reasoning modes is possible. To explain this, we start from a paper by Pothos (2005) who argued that rule- and similarity-based processes can be viewed as opposite ends on a continuum. In his view, a process is similarity based if it is based on multiple features; and it is rule based if it is based on only a small set of features. For example, classifying a particular stimulus as a car based on ten of its features would constitute similarity-based processing; but classifying it based on only one would be rule-based processing. As suggested by some of the commentators to Pothos’ paper (e.g., Smith, 2005), there may be many more dimensions to distinguish rule- from similarity-based processing. Yet, Pothos’ core contribution remains valid in our view, namely that rule- and similarity-based processes are qualitatively similar and that they are only quantitatively different (e.g., in terms of the number of features over which they operate).

Aside from the number of features, it is also possible that the type of features on which the reasoning processes operate distinguishes rule- from similarity-based processing. We propose that one characteristic feature that could distinguish rule- from similarity-based processing is the abstractness of the feature (for other characteristics, see Smith, Langston, & Nisbett, 1992). Consider the example in the Introduction where a child categorizes a dolphin as either fish or mammal. In the former case, he/she may be led by concrete external features (e.g., shape, color, habitat), and in this sense respond in a similarity-based manner; in the latter case, he is probably led by the feature ‘‘nurses its offspring,’’ leading him to respond that the dolphin is a mammal. The feature of a dolphin nursing offspring is usually not visible, and in this sense more abstract because the participant has to generate it himself, but formally the decision process can be conceptualized in a similar way in the two cases.

Applying this line of argumentation to the Shanks and Darby experiment, we argue that the similarity-based participants respond directly to the features presented by the experimenter, whereas the rule-based participants respond to the (abstract) features generated in response to the stimuli. For example, when shown a compound stimulus IJ (with prior training on I+ and on J+) a similarity-based participant responds on the basis of the outcomes that he has experienced to be associated with the single-element stimuli, hence his
“allergy” response (IJ+). A rule-based participant, however, may have (internally) generated the feature “allergy” (+; associated with the single-element stimuli I and J in the training phase), but this feature may prompt itself to be reversed when a food item is presented in combination with another food item. Indeed, the participant has learned that a combination of two identical features from single-element stimuli (e.g., A+, B+) should be reversed in the compound stimulus (e.g., AB−).

The next question, then, is, Why do some participants operate on some features and other participants on others? One plausible reason involves the fact that external features are easier to operate on: these features are, by definition, explicitly presented and can immediately evoke a response, whereas internal features first have to be generated by the participant before responding to them is possible. Thus, if the task is difficult for a participant, he/she may rely primarily on external features, whereas if the task is less difficult, residual resources are available to generate internal features. Many factors could influence task difficulty. One factor might be learning rate; for a fixed number of trials, participants with a small learning rate may not be well advanced in learning the regularity behind the task, having picked up only the superficial features of the task structure. Such participants would then be classified as similarity-based reasoners. On the other hand, participants with a high learning rate would have advanced further and also picked up the deeper task regularities; they would then base their responses on (self-generated) abstract features and be classified as rule based by the experimenter. It follows that a second relevant factor has to be the total number of trials available to a participant.

If rule based and similarity-based reasoning merely differ in terms of the features that are operated upon, it should in principle be possible to devise a neural network model equipped with a standard learning rule that switches from using one set of features (e.g., external) to another (e.g., internally generated) simply by being provided with a higher learning rate or a larger number of trials. The network should gradually become more rule-like as a result. Exploring this possibility is the purpose of our simulations.

3. Model

We used a neural network model that we proposed earlier for a different cognitive domain (multidigit number naming; Verguts & Fias, 2006) in which the input and output units were redefined. This model is an adaptation of a Simple Recurrent Network (Elman, 1990).

3.1. Coding conventions and model architecture

The model and its connectivity pattern are depicted in Fig. 2. We used localist input coding; each of the stimuli A–P (i.e., the stimuli used in Shanks and Darby’s experiment) was assigned its own input unit (16 input units). The use of localist coding was out of convenience rather than representing a theoretical stance. Presentation of a compound stimulus was implemented by sequential activation of two of the input units (e.g., C and D for
Which unit was activated first was randomly determined. Ten hidden units were used. There were two response units, each corresponding to a particular outcome, allergy (+) or no allergy (–), respectively. The model’s response was read off when all inputs were processed (i.e., one input in single-element stimuli, two in compound stimuli) and was determined by which response unit’s activity was the strongest.

An arrow between two layers in Fig. 2 means that there is full interconnection between all units from the sending layer to the receiving layer, except from the hidden layer to itself, where each unit connects only to itself. Hidden units receive activation from input units (arrow 1), hidden units (arrow 5), and response units (arrow 4). Output units receive activation from input units (arrow 3) and hidden units (arrow 2).

### 3.2. Activation equations and learning rule

Activation of a hidden unit was a sigmoid function of the net input to that unit. In particular,

$$\chi_{i}^{\text{hidden}} = \frac{1}{1 + \exp(-\text{net}_i + b_i)},$$

where net$_i$ is the net input to that hidden unit. The net input is the sum of the activation that the hidden node receives from the input, hidden, and output layers. The parameter $b_i$ is a bias parameter. An equation similar to Eq. (1) holds for an output unit.

On each training trial, one of the single-element or compound stimuli was presented at random. After each training trial, connections and bias parameters were simultaneously updated using a generalization of the backpropagation algorithm developed for recurrent network models (backpropagation-through-time; Rumelhart, Hinton, & Williams, 1986). Each update for a weight from unit $j$ to unit $i$ could be formulated as follows:

$$\Delta w_{ij} = \beta a_j \delta_i,$$

in which $\beta$ is the learning rate, $a_j$ is the activation of the sending unit $j$, and $\delta_i$ is the error at the receiving unit $i$. A cross-entropy measure of error was implemented (Plaut, McClelland, Seidenberg, & Patterson, 1996).
4. Simulation 1: Slow versus fast learning (learning rate manipulation)

4.1. Method

Three learning rates (β in Eq. 2) were implemented (β = 0.001, 0.002, and 0.05). For each learning rate, 200 models (simulated participants) with initial random connections were simulated. There were 100,000 training trials for each simulated model. The impact of the direct pathway (arrow 3 in Fig. 2) was slightly weakened by imposing a small bias implemented as an extra weight decay factor −0.01 × wij to the right-hand side of Eq. (2) for the direct connections from input to output (cf. Plaut et al., 1996). This implements the assumption that people will try not to let the direct bottom-up route dominate responding (this assumption is removed in Simulation 3).

4.2. Results and discussion

4.2.1. Training performance

After training each model, its accuracy was tested once on each of the 18 training patterns. Mean accuracy was 15.8 (SD = 1.4), 17.9 (SD = 0.4), and 18 (SD = 0) for models with learning rate 0.001, 0.002, and 0.05, respectively.

4.2.2. Test performance

Generalization performance of the model is shown in Fig. 3A,B. As can be seen in Fig. 3A, the model with low learning rate (0.001) responds similarity based to the compound stimuli, whereas the model with high learning rate (0.05) responds rule based. The same trend appears for the single-element stimuli (Fig. 3B), although the model never comes to respond rule based in this case. Note that this is also the case in the human data (Fig. 1B).

To investigate how the model switches from similarity- to rule-based reasoning with increasing learning rate, we quantified the impact of both the bottom-up (input → response) route and the top-down (hidden → response) route. For the bottom-up route, we did this by calculating, for all single-element stimuli for which feedback was given during the training phase, the correlation between the bottom-up weight and the corresponding target response (1 or 0). This correlation was calculated for each response unit separately, and the mean correlation across simulations was obtained. Mean correlations for the three learning rates are shown in Fig. 4A. As can be seen, this correlation is stable from the smallest learning rate on.

To investigate the impact of the top-down route, consider a weight from a hidden unit to a response unit and the corresponding weight going in the reverse direction (response unit to hidden unit). To solve the task in a rule-based way, the model could make these two weights of opposite sign. For example, suppose stimulus A activates (via the direct input → response connection) an allergy response (+); because of the reversal rule implemented in the experimental design (AB→), if a second stimulus (B) is then processed, the allergy response should be inhibited. This inhibition is unlikely to come from stimulus B.
itself, however, because B on its own also leads to an allergy response. One way to do it
is to make the allergy response activate a given hidden unit, which itself inhibits the
allergy response. This mechanism allows one to solve the triple A+, B+, and AB−,
and if this is how the model does it, a hidden → response weight and its corresponding
response → hidden weight should be of opposite sign, and the correlation across all pairs
of weights should be negative. This correlation was calculated for the allergy response
unit and the no allergy response unit, and the mean correlation over all simulations was
obtained. These correlations are shown in Fig. 4B; all of these correlations were actually
negative, but the absolute value is depicted so that “up” means “more influence” in
every panel. As can be seen in Fig. 4B, this correlation is stronger (more negative) with
increasing learning rate, suggesting that the model gradually picks up on the reversal rule
of the experimental design.

Fig. 3. Percent allergy predictions on transfer items in the simulation data. Rows 1, 2, and 3 correspond to Simulations 1, 2, and 3, respectively. The X-axis represents learning rate in rows 1 and 3, and number of training trials in row 2.
Combining Fig. 4A and 4B provides insight into why the model responds similarity based with low learning rate but rule based with high learning rate: For slow learners, only the bottom-up route (input → response) has been learned correctly, so responding to single-element stimuli succeeds, but the same response is given to the compound stimuli that were not presented in the training phase (i.e., similarity-based responding). For fast learners, the bottom-up route is also operative, but in addition the top-down route via the hidden units has developed as well, and the combination of both makes rule-based responding possible.

A different way of characterizing the two routes is shown in Fig. 4C,D. In Fig. 4C, we plotted the mean weight in the input → response pathway; and in Fig. 4D, the mean absolute weight in the hidden → response pathway. The weights used in Fig. 4D were negative, but again the absolute value was taken to increase readability of the plot (so “up” again means “more influence”). This shows the same trend as in Fig. 4B that the top-down route becomes more influential with increasing learning rate.

Performance on the transfer compound stimuli was captured quite well in the model; reasoning was similarity based early in training but rule based later on (cf. Fig. 3A). Regarding
the transfer single-element stimuli (K, L, O, P), performance of the model with low learning rate was similarity based (cf. Fig. 3B), whereas human participants with low training phase performance were neither clearly similarity or rule based. In the model with a high learning rate, performance on the transfer single-element stimuli is a mixture of rule- and similarity-based reasoning components which cancel each other out, leading to chance level allergy predictions on the transfer single-element stimuli. The same was true in humans with high training phase performance, suggesting that they apply a similar mixture of rule- and similarity-based reasoning. Indeed, contrary to that suggested by Shanks and Darby (1998, p. 406), these participants cannot be said to be applying a rule such as ‘‘A compound and its elements predict opposite outcomes’’; under this rule, given that performance on the single-element and compound training stimuli is equally high (100% and 94% accuracy, respectively, in the empirical data), there should be no dissociation between single-element versus compound transfer stimuli. Instead, the rules that are used for single-element versus compound transfer stimuli must be at least partly dissociated so that compound transfer stimuli can be more impervious to invasion by similarity-based processing than single-element transfer stimuli. This aspect is captured very well by the model.

5. Simulation 2: Number of trials manipulation

Shanks and Darby (1998) suggested that participants start off as similarity-based reasoners and become rule-based reasoners later on in training. To investigate whether this also happens in the model, we fixed the learning rate in Simulation 2 and varied the number of training trials afforded to different implemented models.

5.1. Method

Number of training trials was varied: 1,000, 3,000, or 100,000 trials were given. Two hundred models were simulated for each number of training trials. The learning rate was set at $\beta = 0.05$.

5.2. Results and discussion

5.2.1. Training performance

Mean accuracy was 14.1 ($SD = 1.1$), 17.4 ($SD = 0.9$), and 18 ($SD = 0$) for 1,000, 3,000, and 100,000 trials, respectively.

5.2.2. Test performance

Fig. 3C,D plots generalization performance as a function of the number of training trials. As can be seen, models with few training trials perform similarity based, and models with more training trials are increasingly more rule based. Fig. 4A–D shows that the reason for this is similar to Simulation 1: the model uses only its bottom-up route early in training but comes to rely increasingly more strongly on the top-down route with more training.
6. Simulation 3: Downsizing the bottom-up route

In the previous two simulations, we added a small bias to the direct route to prevent it from being too influential. To make sure that our results do not critically depend on this weighting of the direct and indirect routes, we report a simulation that did not implement this bias.

6.1. Method

Learning rate was manipulated as in Simulation 1. There was no bias on weights in the input $\rightarrow$ response route.

6.2. Results and discussion

6.2.1. Training performance

Mean accuracy was 15.8 ($SD = 1.4$), 17.9 ($SD = 0.27$), and 18 ($SD = 0$) for learning rate 0.001, 0.002, and 0.05, respectively.

6.2.2. Test performance

Results are plotted in Fig. 3E,F. Results are similar to Simulations 1 and 2, except that the model responds somewhat more similarity based than in Simulations 1 and 2 on the transfer single-element stimuli. Fig. 4 shows that, as in Simulations 1 and 2, the influence of the top-down route increases over time. In contrast to Simulations 1 and 2, here the influence of the bottom-up route also increases (see Fig. 4C), although it increases less steeply than that of the top-down route. Hence, the relative influence of the top-down route increases over time, just like in Simulations 1 and 2.

7. Simulation 4

We have suggested that feeding responses back into the response layer is crucial for rule-based responding. To test this, we ran a simulation using a model with similar computational power, but without this feedback possibility. For this purpose, we used a standard feedforward model with hidden units but no recurrency; if our logic is correct, there should be no rule-based responding in this case.

7.1. Method

Stimuli were presented simultaneously, and processing occurred in one time step (hidden unit activation followed by response unit activation). Otherwise, model parameters were taken from the Simulations reported above (10 hidden units; 200 simulations; learning rate = 0.05; 100,000 training trials; no input $\rightarrow$ response bias).
7.2. Results and discussion

Training pair performance was perfect across all 200 simulations, but performance on the test stimuli remained similarity based, with allergy predictions for IJ (100%), O (95%), and P (92%), but not for K (7%), L (8%), or MN (0%). Hence, without recurrency the model did not achieve rule-based performance.

8. Simulation 5

In Simulation 4, recurrent connections were removed, which slightly decreased the number of connection weights (30 weights less than in Simulations 1–3). Hence, the absence of rule-based transfer in Simulation 4 could have been due to decreased computational power rather than to absence of recurrency. In the present Simulation, we vastly increased the computational power of the model by increasing both the number of hidden units and amount of training time.

8.1. Method

The number of hidden units was set to 20, leading to 414 connection weights altogether (in comparison with 254 weights in Simulations 1–3; 224 weights in Simulation 4). The number of training trials was set to 500,000. Otherwise, parameter settings were identical to those of Simulation 4.

8.2. Results and discussion

Results were very similar to those of Simulation 4. Training pair performance was perfect across all 200 simulations, but performance on the test stimuli remained similarity based, with allergy predictions for IJ (100%), O (97%), and P (95%), but not for K (6%), L (4.5%), or MN (0%). This shows that the absence of rule-based performance in models without recurrency is not due to lack of computational power. Instead, because the models of Simulations 4 and 5 do not generate responses sequentially, they cannot use their own responses as input to the response level, and hence abstract rules such as ‘‘if the simple stimulus predicts no allergy/allergy, then the compound stimulus does/does not’’ cannot be detected. As a result, these models do not generalize in the same way as humans or sequential models do. Strictly speaking, it is not recurrency per se that is crucial here; in theory it is possible to construct a feedforward network with the same functionality as a recurrent network (Rumelhart, Hinton, & Williams, 1986). However, this feedforward network would then have much of the ‘‘feel’’ of a recurrent network, as it would use its own response to (part of) the input as a new feature on which to base its final response. As discussed below, it is recycling of responses which is crucial here, and this can most naturally be captured in a recurrent network.
9. General discussion

We have shown that a single model can account for both rule- and similarity-based processing. Depending on learning rate (Simulations 1 and 3) and number of trials (Simulation 2), the model’s reasoning was found to be relatively more similarity or rule based. A critical aspect of the model was the ability to allow its own responses to be used as input for later responses (Simulations 4 and 5). We envision that in general there are also other parameters determining reasoning mode, such as number of hidden units, bias toward a particular route, and so on. Each individual presumably has his or her own parameter configuration, leading to individual differences between participants at the behavioral level (Garlick, 2002). An immediate prediction of the model is that when an experimental task becomes more difficult (e.g., time constraints, working memory load), the more difficult (top-down) route (which learns more slowly; see Fig. 4) will be hindered to a larger degree than the bottom-up route, leading to more similarity-based responding. This is indeed empirically observed (e.g., Smith & Kemler, 1984; Vandorpe, De Houwer, & Beckers, 2005).

9.1. Internal and external features

Of course, the demonstration that rules and similarity can comfortably fit into a unified framework is not new as such and is in fact an important theme in the neural network literature (e.g., Erickson & Kruschke, 1998; Johansen & Palmeri, 2002; McClelland & Rumelhart, 1987; Rougier, Noelle, Braver, Cohen, & O’Reilly, 2005; Rumelhart & McClelland, 1987). Also recurrent models have often been used for modeling (sequential) rule-based behavior in neural networks, in particular based on Elman’s (1990) Simple Recurrent Network model (e.g., Servan-Schreiber, Cleeremans, & McClelland, 1991). Further, many models have also considered the role of internal features in addition to external ones (e.g., Erickson & Kruschke, 1998; Kurtz, 2007; Love, Medin, and Gureckis, 2004; Plaut et al., 1996; Seidenberg & McClelland, 1989; St. John & McClelland, 1990). However, the type of internal features considered in the present paper is significantly different from those discussed in earlier work. The most important aspect is that the internal features considered here are actually possible responses to the stimulus input; for example, when stimulus input IJ is given, the network first “responds” + (allergy) and uses this (internally generated) response as a new feature to elicit the second (and correct) response. This is important because responding based on internal responses allows the network to construct “rules” such as “if the simple stimulus predicts no allergy/allergy, then the compound stimulus does/does not.” This allows for broader generalization than is possible with rules formulated based on input features only. Indeed, it is hard to see how a model would be able to solve the Shanks and Darby task if input features only are used; further, adding hidden units is sufficient to solve the training stimuli, but it is not sufficient for spontaneous rule-based performance (cf. Simulations 4 and 5). Besides the Shanks and Darby data, there are many other data sets in the learning psychology literature that are claimed to constitute proof for two qualitatively different learning systems (e.g., Beckers et al., 2006; Blaisdell,
Sawa, Leising, & Waldmann, 2006). At this moment we pursue the possibility of applying our model to these other findings as well.

Related to this issue, it has often been observed that people shift from using rules to using a similarity-based strategy (i.e., exemplars) in classifying experimental stimuli (e.g., Johansen & Palmeri, 2002; Nosofsky, Palmeri, & McKinley, 1994). At first sight this may seem the reverse shift as found in Shanks and Darby’s (1998) experiment. However, in these cases the rules were based on external features (e.g., visual features of rockets in Johansen and Palmeri’s Experiment 1), whereas in the case considered by Shanks and Darby, the rules are based on internal features. A general principle incorporating both empirical patterns may be that participants base their responses on single external features in the beginning; and with additional time, they either combine different features (as in the categorization tasks) or use more abstract, internally generated features (as in the Shanks and Darby task). Hence, the shift may be characterized not as moving from rules to similarity or vice versa, but rather as moving from similarity computed on the basis of one or a few features to similarity computed by combining multiple features; or phrased more generally, from easy classification on the basis of salient features to more accurate classification based on more complex regularities.

9.2. Similarity and rules united

Our claim that rule- and similarity-based processing emanates from a single learning rule does not deny a dual processing architecture with different brain areas involved in rule- versus similarity-based processing. Indeed, in the categorization literature a vast amount of evidence has been accumulated that the two processes are dissociable, with patient work (Smith, Tracy, & Murray, 1993), neuroimaging (e.g., Smith et al., 1998), and behavioral data (Ashby, Ell, & Waldron, 2003; Erickson & Kruschke, 1998; Johansen & Palmeri, 2002). Similarly, in the learning literature different authors have argued for the importance of rule use either in humans only (Penn & Povinelli, 2007) or in animals and humans (Beckers et al., 2006; Blaisdell et al., 2006). In line with this, different behavioral profiles for both processes are predicted by our model. Further, given that the hidden units and input units of our model are dissociable, the model also predicts that different lesions may lead to different impairments (similarity or rule based) and that with increasing time and increasing learning rate different brain areas will be recruited (cf. Fig. 4). Our main claim is that both rule- and similarity-based processing may emerge from application of a single learning rule, and, crucially, that the two systems operate in a qualitatively similar manner.

Our approach was inspired by Pothos (2005) who attempted to delineate rule- versus similarity-based processing on the basis of the number of features used by either type of process with rule-based processing being based on few features and similarity-based processing being based on many features. As pointed out above, this unidimensional characterization of similarity- versus rule-based processing is not without problems and additional dimensions have been proposed. Rule- and similarity-based processing are more appropriately considered as prototype concepts (Smith, 2005) in which a prototypically
rule-based cognitive process is based on few features, gives differential attention to different features, and can be made verbally explicit. To this list, and based on the simulations reported in this paper, we may add the dimension “is based on internally generated features” (or stated otherwise, “is influenced by top-down pathways”). On the other hand, a prototypical similarity-based process is based on many features, divides attention equally over all features, and is not verbally explicit. In analogy with rule-based processing, the dimension “is based on external features” (i.e., “is bottom-up driven”) may be added.

One of the dimensions that (prototypically) distinguishes rule- from similarity-based processing is the extent to which the underlying regularity is verbalized. Our model does not verbalize its procedures, however, and yet some learning conditions led to very strict rule-based performance. This shows that at least in principle rule-based and verbal processing are dissociable, consistent with Smith’s (2005) analysis. Of course, verbal labels may help in forming rule-based reasoning, leading to a correlation between the feature “verbally explicit” and rule-based processing, but this correlation is not perfect. In a similar vein, also rule-based processing may be dissociated from propositional knowledge, which allows removing the uneasy (and contested; Penn & Povinelli, 2007) assumption that rodents are able to apply propositional knowledge (Beckers et al., 2006).

Recently, also Bayesian models have been used to try to integrate rules and similarity (e.g., Tenenbaum, 2000). In fact, the Bayesian approach bears similarity with our own: Despite the differences between the Bayesian and the neural network approach, both are guided by the common goal of optimization. A Bayesian model attempts to make a statistically optimal inference about which properties a given object possesses given current knowledge (e.g., Anderson, 1991); a neural network attempts to minimize an error function such that the correct properties are associated with each object in the training set. In both approaches, rule- and similarity-based reasoning are two manifestations of the same system optimizing its goal function. In addition, both approaches place strong emphasis on the structure of the environment (in experimental contexts, the training set) for understanding cognition. The main difference between the two approaches is that a Bayesian approach eschews time and architectural restrictions as much as possible, focusing on optimality and the structure of the environment; whereas neural networks do take these restrictions into account. Hence, the two approaches are complementary rather than competitive: whereas the Bayesian approach is unrealistic in its optimality assumption, the neural network approach often needs to invoke architectural assumptions that are unmotivated by empirical data. With these remarks in mind, it is interesting to note that a recent paper attempted to model the Shanks and Darby data from a Bayesian point of view (Kemp, Goodman, & Tenenbaum, 2007). However, this model can only simulate fully rule-based performance (i.e., not the differences between groups or the dissociation between transfer single-element and compound stimuli seen in the empirical data). Nevertheless, we are convinced that further work in both traditions will elaborate the connections between the two approaches and the empirical data.
Acknowledgments

This work was supported by grant P6/29 from Interuniversitary Attraction Poles program of the Belgian federal government. We thank Jan De Houwer, Wim Gevers, David Lagnado, Magda Osman, David Shanks, Maarten Speekenbrink, and Gert Storms for useful comments on the research reported in this paper.

References


