ELSEVIER

Cognition 65 (1998) 197–230

COGNITION

Similarity and rules: distinct? exhaustive? empirically distinguishable?

Ulrike Hahn*, Nick Chater

Department of Psychology, University of Warwick, Coventry CV4 7AL, UK

Abstract

The distinction between rule-based and similarity-based processes in cognition is of fundamental importance for cognitive science, and has been the focus of a large body of empirical research. However, intuitive uses of the distinction are subject to theoretical difficulties and their relation to empirical evidence is not clear. We propose a 'core' distinction between rule- and similarity-based processes, in terms of the way representations of stored information are 'matched' with the representation of a novel item. This explication captures the intuitively clear-cut cases of processes of each type, and resolves apparent problems with the rule/similarity distinction. Moreover, it provides a clear target for assessing the psychological and AI literatures. We show that many lines of psychological evidence are less conclusive than sometimes assumed, but suggest that converging lines of evidence may be persuasive. We then argue that the AI literature suggests that approaches which combine rules and similarity are an important new focus for empirical work. © 1998 Elsevier Science B.V.

Keywords: Similarity-based process; Rule-based process

1. Introduction

The contrast between rule- and similarity-based accounts of cognition is central to cognitive science. The two approaches correspond to different research traditions, and the contrast between them is the focus of vigorous empirical and theoretical debate across a wide range of cognitive domains.

The idea that cognition involves following mental *rules* lies at the heart of the classical picture of the cognitive system (Newell and Simon, 1990). Mental rules encode general facts about the world and these facts are applied to specific instances

* Corresponding author. e-mail: u.hahn@warwick.ac.uk

PII S0010-0277(97)00044-9

in cognitive activity. A paradigm example is language: linguistics aims to specify rules which explicate the structure of language, and it is assumed that these rules are mentally represented and applied in language processing. The same pattern is assumed to hold in knowledge of naive physics (Hayes, 1979), arithmetic (Young and O'Shea, 1981), social conventions (Cheng and Holyoak, 1985), and so on. In all these cases, knowledge is stored in collections of rules, which are organized into *theories*. These are assumed to have the same structure as explicitly-described theories in science: collections of general statements from which predictions and explanations for specific aspects of the everyday world can be constructed. The emphasis on rules is embodied in many formalisms used in psychological modeling, most directly in systems based on production rules (Newell and Simon, 1972; Anderson, 1983; Newell, 1991) or on logical inference (Inhelder and Piaget, 1958; Braine, 1978; Rips, 1994). It is also embodied in much practical artificial intelligence research (within what Haugeland calls GOFAI – good old fashioned AI (Haugeland, 1985)) ranging from early game-playing programs (Newell, 1963) to expert systems (Dayal et al., 1993).

Similarity, in conjunction with sets of stored instances, suggests an alternative model of cognition. Instead of deriving general rules concerning the structure of the world, past situations ('instances' in psychology; 'cases' in AI) are stored in a relatively unprocessed form. Reasoning concerning a new situation depends on its similarity to one or more past situations. Here, we shall call such methods similarity-based reasoning, to emphasise the centrality of similarity. Such theories have been proposed in many contexts: in exemplar theories of concepts (Medin and Schaffer, 1978; Nosofsky, 1984), in instance-based models of implicit learning (Berry and Broadbent, 1984, 1988; Vokey and Brooks, 1992; Redington and Chater, 1996), in theories of reasoning (Ross, 1984, 1987; Ross and Kennedy, 1990), in 'case-based reasoning' in AI (Kolodner, 1991; Aamodt and Plaza, 1994), and 'lazy learning' (Aha, 1997) in machine learning. Moreover, similarity-based approaches to cognition are strongly rooted in behaviorist theories of human learning. The fundamental behaviorist claim is that behavior is mediated by a set of stimulus-response associations. As no stimulus is exactly the same as any previously-encountered stimulus, behavior must depend on some form of *generalization*, depending on the similarity between the new stimulus and previous stimuli. Thus, similarity has been stressed by behaviorists in psychology (Pavlov, 1927) and philosophy (Quine, 1960), as well as being important in cognitive science.

The difference between rule- and similarity-based accounts is clearly of central theoretical importance to cognitive science, but can they be distinguished empirically? There have been many attempts to do so in areas of cognitive psychology as diverse as categorization (Komatsu, 1992), implicit learning (Reber, 1989; Shanks and John, 1994), problem-solving (Gentner, 1989) and the development of reading skills (Goswami and Bryant, 1990). Our own research has concentrated on empirically distinguishing specific rule-based accounts from similarity-based (and other) accounts in the context of language and implicit learning (Redington and Chater, 1994, 1996; Nakisa and Hahn, 1996; Hahn et al., 1997; Nakisa et al., 1998).

However, the interpretation of the empirical evidence is difficult, because both classes of account are very heterogeneous, and hence each can capture a wide range of data. Possibly, these two classes are too broad to really allow an overall empirical assessment. Perhaps, the best empirical research can do is to test particular models of each kind, not 'rules' or 'similarity' generally. Furthermore, the empirical literature contains many confusingly correlated distinctions such as symbolic versus subsymbolic, abstract versus specific or deductive versus inductive. The resulting problems of interpretation which we have encountered in our own work have acted as a personal motivation for developing the ideas in this paper: specifically, to provide a core conceptual distinction between rule- and similarity-based models as a framework for interpreting empirical and computational considerations.

One reaction to the difficulties encountered in distinguishing between such broad classes of account is simply to abandon the attempt. However, we believe that this is at best a last resort: there are strong reasons to attempt to maintain the rule- versus similarity-based distinction. If viable, it allows general theoretical statements to be made about broad classes of account, which are not tied to specific models or implementations. This is important for relating different theoretical proposals, and for unifying what must otherwise remain fractionated literatures on different cognitive domains. Moreover, the viability of a general distinction has been routinely presupposed by the empirical literature, wherever the terms 'rule-' or 'similarity-based' are used without further clarification, such as in the many experimental studies which seek to distinguish rule- and similarity-based reasoning on the basis of their putative effects, without further commitment to more specific models (e.g. Reber, 1989; Shanks, 1995). The issue of whether a general distinction can be maintained is therefore clearly a pressing one. This paper has three main sections. In Section 2, we show how both classes of account appear so general that they seem to collapse into each other. Section 3 provides a core account which successfully separates rule- and similarity-based processes. We show that this core account correctly decides intuitively clear cases of each type more adequately than a range of alternative criteria that might be suggested. In Section 4, we re-assess how empirical and computational evidence can be brought to bear on distinguishing between the two classes of model and, finally, consider implications for future research.

2. Rules and similarity: the problem

The intuitive notions of both rule- and similarity-based processes seem alarmingly general[1]. Almost any aspect of thought may be viewed as determined by

[1]Moreover, the very notions of 'rule' and 'similarity' have been attacked in the philosophical literature (Goodman, 1972; Kripke, 1982). However, these problems are so general that they threaten the entire program of cognitive science, rather than providing specific difficulties for the present debate (Hahn, 1996; Hahn and Chater, 1997).

rules, at least in the sense that laws of nature are rules of a kind; and similarity seems an essential ingredient of an extremely wide range of paradigms and phenomena – connectionism, case-based reasoning, exemplar- and prototype-theories, and possibly even metaphor and analogy.

The threat that follows from the generality of both 'rule' and 'similarity' can be illustrated by the apparent possibility of each account 'mimicking' the other.

First, as suggested by Nosofsky et al. (1989), 'rule' can be used to include procedures for computing similarity as special cases. Indeed, specific theories of similarity, such as geometric models (Shepard, 1980) or the contrast model (Tversky, 1977) appear to provide suggestions about what this rule might be.

Second, 'similarity' appears so general that it can include any rule. Suppose we view a rule, R, as a function from inputs to outputs. Define a dissimilarity measure, *D*, such that

$$D(x, y) = 0 \textit{ iff } R(x) = R(y)$$

$$D(x, y) = 1 \textit{ otherwise}$$

That is, two inputs are similar when the rule gives the same output for both and dissimilar otherwise.

Therefore, similarity-based reasoning might be viewed as involving a kind of rule; and rule-based reasoning might be viewed as involving a kind of similarity. The notions seem so general that they collapse into each other.

The artificiality of this 'mimicry argument' may lead one to underestimate the extent of the problem. However, more realistic variants abound. Allen and Brooks (1991) discuss 'additive rules of thumb' of the form '*At least two of* (*long legs*, *angular body*, *spots*) *then* **builder**.' These rules, however, are equivalent to a special case of a psychological prototype model (Smith and Medin, 1981) – where the prototype is defined by *n* features, of which *m* must be present – which seemingly involves similarity comparison of the new item with the prototype. Moreover, the same behavior can be obtained from a single-layer connectionist network with a linear threshold unit. Therefore, identical behavior appears consistent with rules and similarity, as well as with connectionist networks.

Connectionist networks themselves further illustrate the problem, in that they might be seen to fall in both camps. Back-propagation networks are often described as depending on similarity (Rumelhart and Todd, 1993). However, they are also often described as using 'implicit rules' which can be extracted using appropriate analysis (Bates and Elman, 1993; Hadley, 1993; Andrews et al., 1995; Davies, 1995). Therefore, back-propagation networks appear rule- *and* similarity-based.

These concerns suggest that the intuitively sharp distinction between rule- and similarity-based processing may be illusory. If this conclusion is accepted, then the empirical debate aimed at testing between the two is futile. We will argue that this pessimistic conclusion is not justified, that a core distinction can be made, and that empirical evidence, both from experimental and computational sources, can be brought to bear on whether specific cognitive processes are similarity-based, rule-based, or neither.

3. Rules versus similarity: an explication

An explication of the core distinction between rules and similarity must balance two forces. It must be sufficiently specific that it solves the problems of generality that we have outlined. However, it must also be sufficiently general to take in the great diversity within each type of account. Thus, rule-based processes may invoke symbolic statements with logical connectives, with or without explicit variables (as in classical AI, or some parts of the psychology literature (Nosofsky et al., 1989; Sloman, 1996)); they may operate over banks of connectionist units (Touretzky and Hinton, 1988) or have the form of the additive rules of thumb (Allen and Brooks, 1991) mentioned above. Equally, similarity-based models range from case-based reasoning (CBR) systems in AI, where similarity is assessed between graph structures (Branting, 1991), to spatial and set-theroretic models in psychology where similarity is defined in terms of spatial distance or feature overlap, respectively (Shepard, 1957; Tversky, 1977).

One approach to constructing a core distinction proposes that the two classes can be distinguished because they use *different types of representation*. Perhaps rules contain variables but things entering into similarity comparisons do not; or rules are *general* whereas similarity-based reasoning applies to specific claims (e.g. describing specific instances)[2]; or rules are rigid, whereas representations used in similarity comparison are in some sense fuzzy. Whether explicitly or implicitly, such criteria underlie many definitions of rule-following in cognitive science (e.g. Sloman, 1996).

This focus on different types of representation is undermined by the fact that the *very same representation* can be used both in rule- and similarity-based processing. Consider a representation of the information that monkeys like bananas. This can be used as a rule, on encountering a particular monkey, and classifying it as liking bananas. However, it can also be used in similarity-based reasoning in proposing the generalization that gorillas also like bananas. The core distinction cannot simply be based on different *types* of representation; rather, it must involve the way in which representations are *used*.

To clarify, let us consider a specific scenario. Suppose that we are presented with a new item, which is represented by the features {*large*, *barks*, *brown*, *furry*, *has-teeth*... }. To classify this item, we must somehow relate its representation to our existing knowledge. Rule- and similarity-based processes differ regarding the way the representation of the new item is integrated with existing knowledge.

A paradigmatic case of rule-based processing runs as follows. Existing knowledge is stored in conditional rules (e.g. 'if something barks and is furry, then it is a dog'). If the antecedent of a rule is satisfied (it barks and is furry), then the category in the consequent applies (it is a dog). A paradigm case of similarity-based processing is as

[2]From a logical point of view, a natural fromulation of this type of claim is that rules involve *universal* quantification, whereas similarity is defined over instances which are represented by *existential* quantification. Aside from the difficulty pointed out below, this approach collapses because of the purely logical result that any representation involving existential quantification can be converted into a sentence involving universal quantification, and vice versa, by applying negation.

follows. Knowledge is stored as a set of past instances, with associated category labels. The new item is classified as a dog if the past instance to which it is most similar was classified as a dog. In both paradigms, there is a 'match' between the representation of the new item and a representation of stored knowledge. Crucially, however, the nature of the matching process differs in two ways.

First, the antecedent of the rule must be *strictly* matched, whereas in the similarity comparison matching may be *partial*. In strict matching, the condition of the rule is either satisfied or not – no intermediate value is allowed. Partial matching, in contrast, is a matter of degree – correspondence between representations of novel and stored items can be greater or less. Notice that there is no restriction on the nature of the representation that is matched, whether strictly or partially. Our example is implicitly a conjunction (the item must be furry *and* a dog for the rule to be satisfied), but the condition of a rule could equally well be disjunction (furry *or* a dog), or have any form whatever.

Second, the rule matches a representation of an instance (large, barks, brown, furry, has-teeth...) with a *more abstract representation* of the antecedent of the rule (barks, furry), whereas the similarity paradigm matches *equally specific* representations of new and past items. The antecedent 'abstracts away' from the details of the particular instance, focusing on a few key properties.

Note, crucially, that abstraction here is *relative*, not absolute. Thus, a similarity-based process could operate over highly abstract representations, such as logical forms of sentences. Rule-based processes can apply to arbitrarily specific representations (e.g. specifying minute detail about the perceptual properties of the objects it applies to) if the representations of new objects are even more detailed. Thus evidence for highly abstract mental representations is not thereby evidence for rule-based processing; and evidence for highly specific mental representations is not evidence for similarity-based processing. This point will be important in our reevaluation of empirical criteria for distinguishing between rule- and similarity-based processes below.

The need for relative abstraction stems from its link with generalization. If the antecedent of a rule were as specific as the representation of the instance, then it would apply to at most this single instance and thus provide no basis for generalizing our knowledge about one case to another. Indeed. a system containing rules of this form is simply a 'memory bank' of instances and their classifications.

Thus our paradigm cases differ along two dimensions, defining a space of possibilities illustrated in Fig. 1.

This space provides a useful framework for differentiating 'rules' and 'similarity' because it allows us to think generally about the different ways in which stored knowledge can be applied to enable the processing of novel items. The paradigm case of processing corresponds to the top right corners 'strict matching/abstraction.' The paradigm case of similarity-based reasoning corresponds to the bottom left corner, 'partial matching/no-abstraction.' Note that these locations are sufficiently general to be compatible with the diverse array of specific instantiations of 'rule' and 'similarity' mentioned at the beginning of this section: strict matching to an abstraction is not a notion which refers to particular rule formats, nor does partial matching

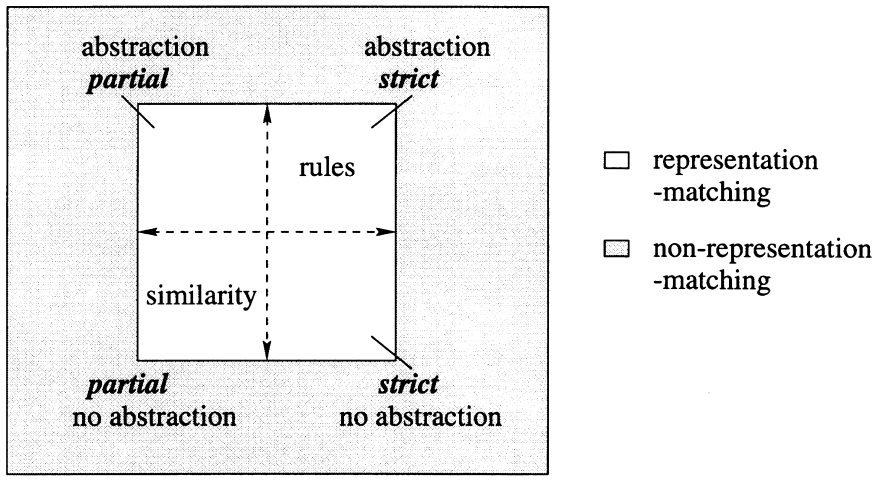

Fig. 1. The space of possibilities for representation matching.

to an instance distinguish between the matching of graph structures, feature overlap or occupying a nearby point in similarity space.

The bottom right corner, 'strict-matching/no-abstraction' corresponds to the 'memory bank.' What about the final corner, at the top left, combining partiality and matching to an abstraction? A candidate for this class are versions of prototype theory, which relax the classic definitional account: i.e. prototypes are construed as lists of core features (hence abstraction) of which only sufficiently many, but not all, must be matched by a new instances (see Komatsu, 1992), thus replacing strict with partial matching. This raises the possibility, which we discuss below, that similarity may encroach into this corner of the space.

Outside the space entirely are processes which do not involve matching the novel instance to stored representations of any kind and, hence, are alternatives to both rule- and similarity-based accounts. For example, generalization might be based on simple failure to discriminate different perceptual stimuli, rather than on stored knowledge. More interestingly, an input-output mapping might be performed without consulting *any* stored representations.

3.1. Why representations matter

We have explained rule- and similarity-based processing in terms of matching between *representations*: of the new item and of stored knowledge. Crucially, it is not sufficient for a process to behave 'as if' it were matching representations. We now illustrate why this is so, considering rules and similarity in turn.

Regarding rules, the issue is the vital distinction between *rule following* and merely *rule describable* behavior (Chomsky, 1980, 1986; Searle, 1980; Dreyfus, 1992; Smith et al., 1992; Marcus et al., 1995). Rule-based reasoning implies *rule-following*: that a *representation* of a rule causally affects the behavior

of the system, and is not merely an apt summary description[3]. Thus, only claims about rule-following are claims about *cognitive architecture*. To illustrate with the classic example, the planets exhibit rule-describable behavior, concisely predicted by Newton's laws, but the planets do not rely on mental representations of Newton's laws to determine their orbits; thus, they are not rule-following. By contrast, explaining why a motorist obeys traffic regulations makes reference to mental states, i.e. knowledge of these regulations. Hence, the behavior in question exhibits rule-following. As *any* regularity can be stated in a format which fulfils our intuitions about 'rule', (e.g. as a universally quantified statement) *any* regular behavior would be 'rule-based' if the distinction between rule-following and merely rule-describable were not maintained; the notion of rule-based processing would collapse into triviality.

Analogous considerations apply to similarity. Unless similarity-based models involve comparison between representations, then *any generalization* (rule-based, similarity-based or even non-representational) can be viewed as similarity-based in the sense that items to which generalization applies can, by virtue of this fact, be viewed as 'similar.' However, such post hoc measures of similarity have no explanatory value. If the constraint of representation-matching is relaxed, the splashes of similar rocks thrown into water could be viewed as similarity-based processing on part of the water, given that they cause similar splashes.

For the rule versus similarity debate to be meaningful, matching must apply to actual representations of rules and instances. Consequently, non-representational approaches to cognition, such as situated robotics (Brooks, 1991), stand outside this debate altogether. Furthermore, *mere procedures* cannot constitute rule-based reasoning. Some confusion over this exists with respect to inference rules in cognition, such as *modus ponens*. Smith et al. (1992), for instance, distinguish rule-following and -describable behavior (they call the latter 'conforming' to a rule) and state that they are only concerned with the former (p. 3). When it comes to inference rules, however, they credit a system with rule following, albeit of 'implicit rules', even if a rule is 'only implemented in the hardware and is essentially a description of how some built in processor works' (p. 34). However, for modus ponens to be *followed*, it is not sufficient for modus ponens to be 'built in' to the procedures by which the system operates. Such a proceduralized notion of modus ponens is found in production rule systems. Production rules 'fire' when their antecedent is satisfied and produce a consequent; however, there is no *representation* of modus ponens. In the rule-following sense, modus ponens *itself* is not a rule in a production rule system, any more than the planets implement Newtonian mechanics. If a proceduralized notion of rule and rule-following is allowed, then the distinction between rule-following and rule-describable behavior is lost again, with the consequences outlined above. As elsewhere, the central question of whether human thought can be described by logical rules or norms must be carefully distinguished from the issue of how such inference is realized in the cognitive system.

[3]This also means that the philosophical debate on rule-following is of direct relevance here (Kripke, 1982; McDowell, 1984; Collins, 1992; Ginet, 1992).

Finally, these considerations also allow us to clarify the nature of standard back-propagation networks, which, we noted above, are claimed both as rule- and similarity-based. On our analysis these networks neither compute similarity nor apply rules, because they do not involve matching to a stored representation of any kind. What representations could be held to be 'matched' with the input pattern? Past inputs are not stored, so that instance-based comparison seems ruled out. The only candidate appears to be weight vectors, but these are not *matched*, i.e. brought into correspondence with, the input at all. Instead, activation flows through the network as a complex non-linear function of inputs and weights[4].

That the network's behavior can be *described* with rules and that the regularities it uses may be 'extracted' (Andrews et al., 1995) is not to say that the network itself is following rules. Likewise. it is true that networks to some extent *depend* on similarity (Rumelhart and Todd, 1993); similar inputs will tend to produce similar outputs. This, however, is a causal story, due to similarity between inputs in the sense of 'overlap of input representations' and, thus, similar activation flow through the network. It is not due to the fact that similarity is being computed, any more than similar rocks producing similar splashes results from computation of similarity.

In summary, for the debate between rule- and similarity-based accounts to be meaningful, matching must apply to *representations* of rules or instances. Thus important classes of cognitive architecture in which no matching to representations takes place stand outside the rule- versus similarity-based processing debate entirely.

3.2. Exploring representation matching: are rules and similarity exhaustive?

We have outlined a core account of the distinction between rule- and similarity-based processing. We now consider some ways in which these notions may be made more specific, and also whether there are other styles of processing, distinct from rule- or similarity-based processing within the representation matching framework. Leaving aside the 'memory bank' which does not generalize to novel items at all, we consider each of the three non-trivial regions of our space – indicated in Fig. 2 – in turn probing the exhaustiveness of 'rule' and 'similarity'. This analysis also shows why our core distinction does not succumb to the mimicry arguments, which appear to collapse rule- and similarity-based processing.

[4]The *inner product* between the input- and the first layer of weights can viewed as a measure of similarity (Jordan, 1986), but only if input vectors are of standard length. If not, our basic intuitions on similarity (Section 3.2.1.) are violated: (1) similarity is not maximal in the case of identity, (2) input vectors – viewed as points in a multi-dimensional input 'feature'-space – which are more distant, i.e. have fewer properties overlapping with the weight vector, can have larger inner products than nearby input vectors due to the effects of length. While normalization is used in some connectionist architectu es such as self-organizing networks (Rumelhart and Zipser, 1985) – and here it may be useful to think of the weight vector is representing a prototypical instance in input space – it is generally not true for back-propagation networks. Even less can we see weight vectors representing rules, and mere procedures, on our account. do not suffice.

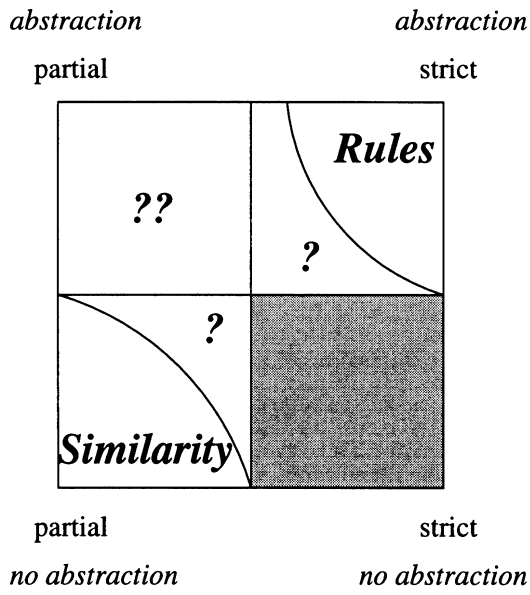

Fig. 2. How far do rules and similarity exhaust representation?

3.2.1. Partial matching and no abstraction: relation to similarity

We have argued that paradigmatic similarity-based processes involve partial matching with no abstraction. However, does similarity matching have additional crucial features and hence are there other forms of matching of partial, non-abstract matching? There are many positive examples of measures of similarity, including the geometric (Shepard, 1980) and contrast (Tversky, 1977) models prominent in psychology and a host of measures in the computational literature on instance-based approaches such as case-based reasoning (CBR) (Herbig and Wess, 1992) and nearest neighbor algorithms (Cost and Salzberg, 1993) in machine learning. However, when it comes to delimiting what exactly counts as 'similarity', our underlying intuitions seem remarkably vague. Indeed, they appear to be exhausted by the following criteria:

1. similarity is some function of common properties[5]
2. similarity is graded
3. similarity is maximal for identity

Thus, any function from common properties to a value on a multi-value scale, which is maximal for identity, will fit the bill.

This is vague, but specific enough to defeat one half of the mimicry argument,

[5]Where 'property' covers binary attributes, continuous valued dimensions and relations.

because our function *D* is ruled out: 'common properties' degenerate to one; the function is not graded; nor is it really 'maximal for identity' because, although it returns the maximal value for identical instances, it also returns this value for any other instance which is an instance of the rule.

Are there other types of partial matching with no abstraction, which are not based on similarity? Possible alternatives arise from considering other information which might be included in the matching process in addition to 'common properties'. Information such as frequency or recency might usefully be used in a function performing partial matching, but these do not seem to be aspects of similarity. Hence, it seem that a whole range of 'hybrid' functions is conceivable. Furthermore, to the extent that our vague intuitions (Goodman, 1972; Goldstone, 1994a) about similarity are made more specific, going beyond 'partial matching without abstraction' this automatically means that matching functions in the bottom left corner which do not meet these criteria will not be similarity functions. So it seems that, whatever theory of similarity is ultimately adopted, there will be other types of 'partial matching without abstraction' not classified as involving similarity.

3.2.2. Strict matching with abstraction: relation to rules

Rule-application, we have argued, is a matter of strict matching to knowledge which is more abstract than the new item to be matched. By definition, this means that the set of objects consistent with the representation of the new instance is a proper subset of the set of objects consistent with the antecedent of a rule.

How do we know that the stored information is more abstractly represented than the new item? In some representation languages, different 'types' of representation reflect different levels of abstraction. Many representation schemes make no such overt distinctions, however. Indeed, natural language makes no surface syntactic distinction between the statements: 'the dinosaur is extinct' from 'the dinosaur is in the museum.' Yet the former can be used as a general rule when classifying newly-encountered animals, whereas the latter describes a specific event and cannot be applied to new instances of any kind.

In considering different representational schemes, from natural language statements to templates with slots and fillers, semantic networks and production rules, two ways emerge in which the general idea of abstraction is manifest. First, a representation can be more abstract than another by 'underspecification': the more abstract representation simply specifies fewer constraints. This is exemplified, for instance, by the use of variables in production system rules (Anderson, 1983), or unification of feature structures in computational linguistics (Shieber, 1986). It is also present in our 'dog-rule', above, where only 'furriness' and 'barking' matter, and other properties are irrelevant. The second way in which a representation can display greater abstraction is through the use of 'general terms.' This implies a hierarchy of terms (e.g. dog, mammal, animal). A description is more abstract if it contains predicates of which the predicates of the less abstract description are proper subsets. This relationship would hold between our 'dog-rule' and instance descriptions phrased in terms of 'short-fur', 'long-fur' etc. The example also illus-

trates that underspecification and general terms can be, and frequently are, combined[6].

Therefore, there are many ways in which an internal representation might be more abstract than the instance-representation with which it is matched, but what type of internal representation counts as a rule? Artificial intelligence and cognitive psychology offer a wide range of models for internal representation, from declarative statements in Prolog, through semantic networks, property list, feature vectors to symbolic systems implemented in connectionist hardware. Which of these constitute 'rules'? Must these be propositional or expressed in a language, possibly encompassing symbols and logical connectives? Adopting any such further constraints on the notion of rule restricts the scope of the notion within the top right corner of the space (Fig. 2), making rule-based reasoning non-exhaustive even of strict matching to an abstraction.

Again, however, even the core notion successfully deals with the second half of the mimicry argument, i.e. that similarity comparison is rule-based because any similarity metric can be specified as a rule. First, it is an empirical question whether the similarity metric is in fact *represented* as a rule in specific cognitive processes. although computational systems can contain an explicit representation of their similarity metric, this metric can equally be proceduralized, just as modus ponens can (see e.g. Kruschke's (1992) implementation of Nosofsky's (1988) generalized context model). If the similarity metric *is* explicitly represented, the similarity comparison, strictly speaking, does involve rule-application (of the metric). However, the rule the system is applying is then so general that it is neither an interesting nor useful claim to say that the system is 'rule-based.' In particular, this claim is not the one that cognitive science is concerned with, because it concerns how the matching process itself is implemented, not the crucial issue of what type of representation-matching is used.

How does the core distinction deal with more realistic examples of mimicry? Let us reconsider Allen and Brook's 'additive rule' prototype models and the equivalent connectionist network. Allen and Brook's 'additive rule' *is* a rule, according to the core distinction, because it requires strict matching of its antecedent (i.e. it applies just when at least the specified number of criteria are fulfilled). However, comparison with a prototype involves similarity (assuming that similarity is well-defined between prototypes and exemplars – see below) and hence this model is similarity-based. Although the two processes produce identical results, they involve different kinds of matching to different representations (a declarative specification that *m* of *n* features must be fulfilled vs. a prototype). Finally, the equivalent single-layer connectionist network does not involve any kind of matching to stored knowledge and hence it is neither rule- nor similarity-based. Thus, the core distinction preserves the

[6]Our notion of abstraction requires *some* loss of information relative to a corresponding specific representation. Hence, we reject the notion of 'ideal abstraction' whcih retains all information (Barsalou, 1990). In fact, information loss is present even in Barsalou's examples where the abstractions contain all the properties of the exemplars but 'centralized'; the centralized, 'abstract', representation no longer contains sufficient information to reconstruct the particular exemplars. As noted, an 'absraction' which retained *all* information about a set of instances could not be used in generalization.

intuitive sense that the three models achieve the same result in very different ways.

The fact that rule- and similarity-based processes can produce equivalent classifications may seem to undermine the empirical testability and even the theoretical importance of the distinction. It is important to stress, however, that these processes do differ in a wide variety of cognitively important 'secondary properties' e.g. in the learning procedures required for acquisition, ease with which modification can be affected, or behavior under noise (see also Hahn, 1996) Thus, the rule/similarity distinction is important for cognitive science, although not always for primary input-output behavior.

In summary, the conclusions on the scope of rules parallel those on similarity. On the one hand, new ways of achieving 'strict matching to an abstraction' may emerge; on the other hand, the notion of rule might be tightened up by adopting further constraints. Consequently, it seems unlikely that 'rules' will ultimately exhaust the space of strict matching with abstraction.

3.2.3. Partial matching to an abstraction

This leaves the 'top-left corner': partial matching to an abstraction. Rules do not seem to spill over into this corner; neither legal rules, physical laws, universally quantified formulae in first order logic, probabilistic rules, nor defeasible rules allow partial matching. However, similarity might extend into this corner, specifically in those versions of prototype theory in which the prototype is an abstraction of typical or core properties that, as neither necessary nor sufficient, need be matched only partially.

Can such partial matching to an abstraction count as a similarity comparison? Current theories of similarity differ on this issue. Tversky's contrast model, for example, allows similarity comparison between *any* two featural representations, even an item and its category.

Geometric models of similarity do *not* allow comparison between an instance and something that is an abstraction relative to this instance; instances are points in similarity space, but abstractions over instances are regions therein, and the notion of 'distance' (and thus, similarity) between a point and a region is not defined. Thus it depends on the choice of similarity theory whether similarity-based processing extends to partial matching with an abstraction. It is not the purpose of this paper to decide such issues and the task of empirically distinguishing rules and similarity can be investigated without legislating terminology on this point. In evaluating empirical evidence for rules or similarity, the existence of this part of the representation-matching space must be born in mind. What it is called is of secondary importance, last but not least, because, here too, it is unlikely that similarity would ultimately exhaust partial matching to an abstraction, even where a theory of similarity allows such matching.

Raised by these issues is the general relationship between similarity- and instance-based reasoning. However, regardless of how the case of partial matching to an abstraction is decided, 'similarity-based' is wider than 'instance-based' if (as is generally the case in psychology (Medin and Schaffer, 1978; Nosofsky, 1988))

'instance' refers only to actually-encountered exemplars. Alternative notions of prototype such as the central tendency or a modal exemplar (Posner and Keele, 1970; Rosch et al., 1976; Komatsu, 1992; Nosofsky, 1992) are straightforward cases of partial matching between representations of the same level of abstraction and thus constitute similarity-based processing for all similarity theories.

3.3. What the distinction is not

We now examine other criteria that might be viewed as relevant to the distinction between rules and similarity. In contrast to our core distinction, these potential alternatives, although relevant and important in their own right, turn out to cross-classify rules and similarity or, at best, to partially correlate with one or the other.

3.3.1. Types of computational architecture

Similarity-based methods are sometimes associated with highly parallel, distributed computational architectures. Rule-based processes, by contrast, are sometimes associated with serial, symbolic computation.

3.3.1.1. Serial versus parallel. The serial-parallel distinction does not distinguish between similarity- and rule-based approaches. Production rule systems, which are paradigm rule-based systems, have both serial and parallel implementations. On the side of similarity, most CBR systems, paradigmatic similarity-based approaches, have serial implementations, although (partially or completely) parallel implementations are possible e.g. Myllymäki and Tirri, 1993; Brown and Filer, 1995.

3.3.1.2. Symbolic versus connectionist. The border between rule- and similarity-based processes also fails to coincide with the distinction between symbolic and connectionist computation. First, 'symbolic' does not equate to 'rule': similarity-based systems such as CBR systems (Aamodt and Plaza, 1994) and nearest neighbor algorithms in machine learning (Aha et al., 1991; Cost and Salzberg, 1993) are typically symbolic. Second, 'connectionist' does not equate to similarity – indeed, we have seen that the most widely used connectionist networks, back-propagation networks, are neither rule- nor similarity-based.

3.3.2. Structured versus non-structured representations

In psychology, similarity-based methods frequently apply to simple representations, such as vectors of binary feature-values (in Tversky's contrast model) or numerical values (in geometric models), whereas rule-based models typically use structured representations, involving arbitrarily complex symbolic expressions. This distinction, however, is also orthogonal to the division between rule- and similarity-based reasoning. Similarity can be defined over structured representations (Gentner and Markman, 1994; Goldstone, 1994b; Hahn and Chater, 1997) – frequently, this similarity relation is called *analogy*, because structural similarities are of central importance (Gentner, 1983; Gentner and Forbus, 1991). Conversely, rule-based

accounts may be sufficiently simple that structured representations are not required (as in most statistical contexts).

3.3.3. Abstract versus concrete representations

In psychology, similarity-based models are often applied to very concrete, typically perceptual, representations of simple stimuli, such as schematic faces or geometric shapes (Reed, 1972; Nosofsky, 1988). More abstract domains, involving reasoning about social interactions (Cosmides, 1989), norms of behavior (Cheng and Holyoak, 1985) or naive physics (Hayes, 1979) are often viewed as involving general rules. We have seen, however, that the rule/similarity distinction does not relate to absolute level of abstraction. Similarity can operate in highly abstract domains (e.g. using analogy in mathematical or scientific reasoning (Gentner, 1989) or case-based reasoning in law (Ashley, 1990; Aamodt and Plaza, 1994)); and rules may be used to process highly specific representations (as, for example, in 'blackboard' models of perception (Selfridge, 1959)). Crucially, we saw that the same representation in stored knowledge can serve either as a rule or an instance in a similarity comparison, depending on the nature of the matching process with new items.

Thus, psychological evidence concerning the level of abstraction of representations does not thereby provide evidence concerning whether cognitive processes are rule- or similarity-based. We shall see the implications of this in discussing attempts to empirically distinguish the two styles of processing below.[7]

3.3.4. Rigidity and gradedness of classification

The issue of rigid versus graded classification may appear to map on to the distinction between strict versus partial matching. Strict versus partial matching, however, concerns how a new *input* is related to existing knowledge; it does not relate to the gradedness of an *output* such as a classification decision. These two issues are distinct, although connected. Specifically, if we assume that the output is a function of the input, then a strict match to an input implies a rigid output: thus rule-based processes produce rigid, all-or-none classifications. However, a partial match to the input is compatible with both a graded and a rigid output, depending on whether information about degree of match is preserved. Degree of match could be used, for instance, to establish graded category memberships. However, other similarity-based classifiers produce yes/no decisions by using a threshold, or competition between instances. For example, the strict nearest neighbor criterion, the paradigm of similarity-based models, bases the decision on the single most similar known item. Thus, the core distinction is related to, but different from, the distinction between processes producing rigid versus graded classifications.

3.3.5. Deductive versus non-deductive reasoning

The distinction between rule- versus similarity-based reasoning also does not map onto the distinction between deductive (certain) and non-deductive inference. While

[7]Evidence concerning abstraction can, of course, be crucially important in distinguishing between specific theories – either kind, which postulate particular levels of representation.

it is true that similarity-based reasoning is never certain, and, hence, always non-deductive, we find deductive reasoning which is not rule-based and rule-based reasoning which is non-deductive.

Probabilistic rules, as well as the non-monotonic or defeasible inference (see e.g. Ginsberg, 1987) necessary to capture how we actually reason with rules such as 'birds fly,' in the face of countless exceptions such as penguins, broken wings and so on, are not deductive (at least in psychological parlance, see Johnson-Laird and Byrne, 1991; Chater and Oaksford, 1996).

We can also find deductive reasoning which is not rule-based, however. '*Or*-introduction,' for instance, allows the inference from $P(a)$ to $P(a)$ or $Q(a)$. Similarly, we can infer from $P(a)$ that $Exists(x)P(x)$. In either case, such an inference constitutes a case of rule-base reasoning only if the 'inference rule' ('or-introduction') itself is explicitly represented and applied (see Section 3.1.), rather than implemented procedurally.

3.4. Summary

We have explicated the 'core' distinction between rule- and similarity-based processing and argued that this explication is distinct from, and more appropriate than, a range of alternatives. We have also shown that our explication is specific enough to clarify unclear cases such as back-propagation networks, and to block the mimicry argument. Having provided this support for our analysis, we now consider its implications for the problem of distinguishing rule- and similarity-based processes empirically.

4. Re-evaluating the empirical evidence

Empirically distinguishing similarity- and rule-based psychological theories has proved to be extremely difficult. Moreover, there are at least two interpretations of the question.

- *Class distinguishability*. Can empirical data distinguish between the classes of rule-based and similarity-based theories? In other words, can we distinguish between rule- and similarity-based accounts of a task *in general*.
- *Specific distinguishability*. Given fully elaborated, *specific* similarity- and rule-based theories, can these be distinguished empirically? This last question is, of course, many questions, depending on what the specific theories are.

In the literature, doubts about class distinguishability have variously been voiced (e.g. Barsalou, 1990; Koh and Meyer, 1991), and we therefore give this issue particular attention[8]. Note that if the classes of rule- and similarity-based theories

[8]However, Barsalou's (1990) argument rests crucially on his notion of 'ideal abstraction', which does not involve any information loss – a notion not allowed in our framework, because information loss is constitutive of what it means to be a.n abstraction, as noted above.

cannot be empirically distinguished, we may nonetheless be able to distinguish fully elaborated, *specific* accounts, i.e. we may be restricted to comparative model fitting.

In terms of our analysis, class distinguishability requires finding empirical evidence which locates a cognitive process in relevant regions of our space of possibilities (Fig. 2) (including, recall, non-representational alternatives, as well as different kinds of matching). Where 'rule' or 'similarity' are given more specific elucidations, they correspond to yet smaller sub-spaces. Thus, empirically establishing a more specific notion also requires eliminating the remaining, rival forms of strict matching to an abstraction or partial matching to an instance. For example, where the use of the term 'rule' is restricted to internal representations which are symbolic and contain variables and logical connectives, an additional empirical case for the symbolic nature of the internal representation has to be made. Likewise, where 'similarity' is restricted to a metric of particular functional form, other forms of partial matching must be ruled out. Hence, dealing only with the broad partitions of the space entailed by our core distinction is the easiest version of the rule/similarity class distiguishability problem and we will focus on how empirical evidence relates here.

Again, the problem is that these classes of account are general, and distinguishing general accounts is difficult in any scientific context. This is because relating general accounts to empirical data requires additional auxiliary (and typically uncertain) assumptions. Indeed, in principle, any theory can be made to fit any data set (Putnam, 1974), by postulating appropriate auxiliary assumptions, requiring additional constraints on what assumptions count as plausible in order to rule out any general account.

In psychology, where any specific mental processes lead to behavior only in conjunction with many other processes which are highly complex but unknown, additional hypotheses which save a theory are particularly easy to provide (Pylyshyn, 1984). Moreover, distinguishing between the classes of rule- and similarity-based accounts of cognition is particularly difficult, because the auxiliary assumptions concern the very essence of the explanation – how instances are encoded, what the actual rule is or what similarities are perceived. This is not a problem of having a single free parameter to tie down: almost the entire predictive content of a rule-based theory depends on what exactly the rules are; almost the entire predictive content of a similarity-based theory depends on what the similarities are.

Perceived similarity depends on which properties of objects are selected as relevant, how they are weighted and what metric is used to compare the resulting representations. The number of possible rules that can be entertained is clearly vast – in principle, there are infinitely many rules which would allow perfect performance in any learning task[9]. Even allowing that this set is restricted by representational and memory limitations, the set of possible rules that an agent might entertain will in most cases still be large. This casts doubt on the possibility of directly distinguishing the classes of rule- and similarity-based theories by a single experimental measure.

[9]For example, in a category learning task, there will be infinitely many rules compatible with the finite set of data (labelled objects) – where the rules differ arbitrarily on the infinite number of unseen examples.

We now consider various suggestions concerning how the classes of theories can, despite the apparent difficulties above, be distinguished, focusing first on experimental criteria and then on computational criteria drawn from AI. Together these will also suggest a different emphasis for future research concerning rule- and similarity-based processing.

4.1. Experimental criteria

We are now in a position to survey empirical evidence which aims to distinguish similarity- and rule-based accounts across various cognitive domains. The literature contains a wealth of potential criteria; we have limited our discussion to those experimental criteria which are widely applicable, i.e. not restricted to a particular question or subject domain.[10] We group this evidence under four headings: effects of instances, effects of rules, generalization beyond the capabilities of instance-based models and patterns of breakdown.

4.1.1. Effects of instances

Similarity-based theories typically assume that new items are compared with representations of old items. If so, then some distinction between performance with genuinely new items (where generalization is required) and old items (which need only be 'looked up' in memory) may be expected.

By contrast, if all items are dealt with using a rule, there seems no reason to suppose that new and old items will be classified differently. This is because the general claim embodied in the rule must be applied to all specific instances; the rule contains no information about which items have been seen before. The ensuing criterion of 'no observable difference between old and new items' is used in many studies (Nosofsky et al., 1989; Allen and Brooks, 1991; Smith et al., 1992) and has been suggested as an 'operational definition' of rule-based performance (Herrnstein, 1990; Shanks, 1995).

4.1.1.1. Old-new recognition. Instance-based accounts, in contrast to rule-based accounts, also require that instances be remembered. Thus, they seem to predict that people should be able to distinguish old from new items, and that their pattern of old-new judgements should relate closely to their categorization performance. Relating categorization and old-new judgements in detail requires a unified psychological account of both. An example of this has been provided by Nosofsky (1988), who obtained a good account of experimental data using simple artificial stimuli. Such a unified account gives strong evidence for similarity-based processing; but failure to find such an account is not equally strong evidence against

[10]Thus we do not consider, e.g. arguments for rules based on linguistic analysis (Marcus et al., 1995), or those of the criteria put forth by Smith et al. (1992) which are tailored specifically to identifying rules of inference. We also omit Sloman's S-criterion (Sloman, 1996), which – stemming from a somewhat different interest – uses conflict to establish two cognitive sub-systems, because further, independent evidence is required to identify these as 'rule' or 'similarity' systems. It is only this further evidence, not the S-criterion, which directly applies to our question.

similarity-based processing: old-new discrimination might, for example, draw on a memory store separate from that used in classification. More importantly, the encoding of an instance may be sufficiently abstract that discrimination is not possible (and, of course, *some* abstraction is inevitable, otherwise no item could ever be re-recognized).

4.1.1.2. Manipulations of the instance-space. This approach is mainly used in studies using stimulus sets which pit rule- and similarity-based classification against each other. The interest lies in the classification of 'critical instances' which, although instances of the intended rule, are more similar to instances of another rule or category. Rule-based reasoning seems to suggest that the 'rule' classification should prevail, whereas similarity-based reasoning seems to suggest the opposite. However, sometimes, effects of both rules and similarity are found (Nosofsky et al., 1989; Allen and Brooks, 1991; Vokey and Brooks, 1994).

Generally, a failure to find instance effects is evidence against instance-based reasoning, at least on the experimenter's assumption of instance encoding and instance similarities, and, conversely, instance-effects can be viewed as refuting reasoning with the rules intended by the experimenter. In both cases, however, refutation of one provides only limited support for the other due to the non-exhaustiveness of rule- and similarity-based reasoning. Where instance effects fail, prototype models (both with and without abstraction, see above) must also be ruled out; they provide an important class of alternatives to rule-based accounts in this context, because 'critical instances' by their very construction are peripheral, and, hence, might be misclassified on prototype accounts as well. Where instance effects appear, simple connectionist accounts of the type we have classified as similarity-dependent but not similarity-based, in our usage, must be ruled out. Consequently, the domain most intensely studied with tasks of this kind, artificial grammar learning (see below), has seen a long series of claims and counterclaims (see Redington, 1996 for thorough discussion).

4.1.1.3. Summary. Instance-space manipulations are an effective tool, but the non-exhaustiveness of rules and similarity means that empirical evidence is more powerful in challenging than in supporting, either account. Also, specific assumptions about rules, instances and instance-similarities must be made, so that this criterion does not pertain to entire classes of account. Memory for instances seems indicative only if a 'unified account' succeeds, making it a powerful but demanding tool. Again, specific assumptions about instances and similarities are required.

4.1.2. Effects of rules

4.1.2.4. Rule priming. Throughout cognition, repetition of the same, or a similar, mental process (e.g. recognizing a word or a picture) speeds performance. Smith et al. (1992) suggest that such priming effects might provide evidence concerning the existence of internal rules. Specifically, they suggest that if priming were observed

between two cognitive tasks which share the same rule, but correspond to very different instances, the rule-based view would be favored. Once we recognize that similarity-based models may be defined over abstract representations, and not merely superficial features of the stimulus, however, it is difficult to rule out the class of similarity-based models.

Langston et al. (unpublished data), for example, use conditional sentences which express either permission or obligation (see Cheng and Holyoak, 1985). They argue that performance on Wason (1968) selection task using these two rules is primed if the underlying rule-type is repeated, even though the surface form of the rules is altered. They argue that this provides evidence for permission and obligation rules. However, this important empirical result is equally consistent with the suggestion that instances of conditional sentences have abstract codes, which distinguish permission and obligation.

Another example is syntactic priming (H. Branigan et al., unpublished data), where sentence production or comprehension is primed by previous sentences which related syntactic structure. This is evidence for abstract representation of syntactic information. However, again, this information may be embodied in rules or as abstract information about the stored sentence-instances (e.g. sentences may be stored not as strings of words, but as labelled tree structures)[11]. In both examples, priming provides important evidence for a particular kind of abstract representation, rather than evidence between rules and instances.

4.1.2.5. Rule complexity. If cognition is rule-based, then the *number* of rules involved in a cognitive task may explain task difficulty (Smith et al., 1992). The number of rules depends, of course, on the specific set of rules under consideration; and difficulty (e.g. time) will also depend on how rules are implemented. In the same way, task-difficulty predictions can be obtained from specific similarity-based models (for example, reaction-time predictions from a specific model of categorization, e.g. Lamberts, 1995). However, there do not appear to be task-difficulty predictions associated with the classes of rule- and exemplar-based models. Therefore, task-complexity considerations appear to be important in testing specific rule-based accounts, but not suitable for distinguishing the rule- and similarity-based classes of account.

4.1.2.6. Verbal protocols. If people use rules, it is possible that they may express these rules, or aspects of them, in verbal protocols. Many production-rule theories of problem solving and skill-learning are based on an interactive process of building rule-based models and matching these models to verbal protocols and task performance (Newell and Simon, 1972). Equally, protocols mentioning comparison with instances (e.g. analogical reasoning) might also provide evidence for similarity-based processes. Protocol evidence is potentially very important. Crucially, its strength depends on the degree to which protocols tie up with other experimental measures, thus providing evidence that protocols are a reliable

[11]Of course, a similarity-based approach to language processing may seem implausible for other reasons.

indicator of the cognitive processes under study. Hence, production system computer models of learning a computer programming language based on protocols which provide good empirical data fits thereby provide evidence for rules (Anderson, 1983). However, there are also circumstances where independent evidence indicates that reported rules were not followed (see Nisbett and Wilson, 1977).

4.1.2.7. Summary. In short, putative effects of rules might provide useful, but not decisive tests, between the classes of rule- and similarity-based accounts. Priming effects may indicate that particular abstract information is represented, but not whether that information is represented by rules or instances. Effects of rule complexity depend on the specifics of the rule-based account and do not follow from the class of rule-based accounts. Finally, verbal protocols may be suggestive, but require evidence that protocols are reliable indicators of the underlying cognitive mechanisms under study. All three criteria require specific rules, although for protocols, these arise directly from criterion use.

4.1.3. Patterns of generalization

In favor of rules, it has been argued that several types of generalization seem inexplicable by an instance-based approach.

4.1.3.8. Extrapolation. Extrapolation in so-called 'function learning' experiments (Koh and Meyer, 1991) provides a potentially powerful source of evidence for rule-based behavior (Shanks, 1995). For example, imagine subjects learning to press a button with a duration proportional to the size of stimuli; larger stimuli requiring longer durations. Correct performance outside the range of stimuli seen so far (e.g. a very long button-press for a very large stimulus) is argued to be incompatible with an instance-based model as the response is not that of the closest previously seen instance, but a novel, i.e. longer, response. Behavior seems to depend on the application of a rule specifying the function relating stimulus height and response duration (Shanks, 1995).

Finding that people can generalize by extrapolation in this way (Koh and Meyer, 1991; Delosh, 1993) is an important empirical result and it constitutes strong evidence against similarity-based approaches. Notice, however, that it does not show that rules are being used. For example, a single-layer connectionist network, with one input and linear output unit, could trivially learn to extrapolate from increasing input to increasing output.

4.1.3.9. Transfer. A further approach considers the transfer of information learned in one domain to another. Perhaps the paradigm example is transfer in artificial grammar learning (AGL) (Reber, 1989). In AGL, subjects try to memorize a set of letter strings, without being told that they are generated according to a set of rules (typically a simple finite state grammar). They are then told about the existence of the hidden rules, but not what the rules are, and asked to discriminate new test strings which do or do not conform to these rules. Subjects perform significantly above

chance in this experiment. One hypothesis is that they have implicitly extracted some of the underlying rules used to generate the items ('implicitly' because subjects typically cannot verbally report any rules they have learned). Another is that they are simply judging the similarity of new items to old, which can also lead to above chance performance. The transfer condition seeks to rule out this possibility by using a different vocabulary of letters in the memorization and discrimination phases. The idea is that the new strings are not at all similar to the memorized strings, and hence similarity cannot mediate generalization. Even here, subjects do show (typically small) above chance transfer performance (Dulaney et al., 1984; Redington and Chater, 1994).

One possible alternative to a rule-based account for this phenomenon is that instance-encoding abstracts away from the specific alphabet used in training, so that instances successfully classify transfer items. Again, we stress that evidence for abstract representations in itself is equally consistent with rule- and similarity-based processes. Abstraction in instance-encoding is perfectly possible, the question is only how much abstraction is *plausible*.

A further alternative is that abstraction occurs only when the stored instances are compared with the transfer stimuli, i.e. at transfer (e.g. Brooks and Vokey, 1991). This is tantamount to analogy and models of this kind equal or surpass human transfer performance without reference to rules (Redington and Chater, 1996). The exact relationship between similarity-based reasoning and analogy is controversial (Seifert, 1989). Thus, analogy presents either a version of similarity-based reasoning or a 'third account'.

Finally, attempts have been made to explain transfer with a connectionist network (Altmann et al., 1995) which appears to involve no matching between input and stored knowledge, and hence falls outside both rule- and similarity-based accounts. Thus, it seems that transfer effects may be explained by rule- and by similarity-based processes and by alternatives in neither framework.

4.1.3.10. Reversal. Another source of evidence comes from reversal of a learned response (Shanks, 1995). In a typical experiment, people or animals are initially trained to associate reliably two distinct responses to two sets of stimuli (Sidman and Tailby, 1982; Vaughan, 1988). Then reversal occurs: it is now the other set which demands the particular response. Subjects are trained to stable performance on the 'reversed' contingency, followed by a second reversal; and so on. After a number of such reversals, both animal and human subjects display the ability to shift almost immediately on new reversed trials, extending their behavior from these first instances to the remaining members of the class. Thus, it is claimed, members of each set are treated as an equivalence class.

This satisfies Shanks' instrumental definition of rule as 'no observable difference between performance on trained items (here, of the reversal trial) and old items' (here, the rest of the class) but it is insufficient on our account. This is because the reversal could be happening solely through a switching of responses *at the response level*. That is, subjects realize that they now have to respond the opposite way from before, e.g. 'yes' now means 'no' and vice versa, but which response to choose (and

then reverse) is still determined solely through similarity comparison with past instances. On this account, subjects need only realize that 'responses have gone funny again'; they need not treat the items as belonging to equivalence classes.

4.1.3.11. Summary. The above three criteria all seek to rule out the entire class of instance-based models. The rule models they aim to support have particular rules in mind – i.e. the underlying function, the rules of the underlying grammar, rules describing the equivalence classes – but any rule which delivers the same classification for the data seen will suffice; hence, these experimental criteria can be seen as distinguishing classes of instances from classes of rules. *Transfer* appears consistent with rules, similarity and connectionist alternatives. *Reversal* appears consistent with both rules and similarity, because it can be explained by switching at the response level. More positively, however, *extrapolation* provides strong evidence against similarity-based models, although it is consistent with non-matching models such as neural networks as well as with rules.

4.1.4. Error and patterns of breakdown

Patterns of breakdown and error provide another potentially valuable source of evidence.

4.1.4.12. Memory failure. Suppose that people learn the rule NOT-RED OR TRIANGLE in an artificial concept learning experiment. Later, they are the tested on generalization to new instances. If their memory is incorrect, they might be expected to classify according to: RED OR NOT-TRIANGLE, or NOT-RED AND TRIANGLE. By contrast, errors on a similarity-based view (based on instances) would not be expected to have this global character. Instead, individual past instances might be misremembered, leading to local misclassifications of nearby novel items. Global errors might, however, result if learning had yielded a single prototype. We are not aware that anyone has aimed to make use of this contrast, but it appears to be a potential direction for future research.

4.1.4.13. Neuropsychology. More dramatically, it is possible that neuropsychological patients may exhibit selective preservation of rules, but loss of exceptions. This appears to occur in reading, with some patients appearing to lose exceptions (surface dyslexia, McCarthy and Warrington, 1986) and others losing the 'rules' of spelling to sound correspondence (phonological dyslexia, Funnell, 1983). This 'double dissociation' has been taken as evidence for rules in reading (e.g. Shallice, 1988).

This may be over-interpreting the data. While connectionist modeling has made us aware of the fact that a uniform connectionist architecture may not adequately capture the neuropsychological data for reading, i.e. that we might need *dual route architectures*, it has also alerted us to the fact that possibly *neither* route need contain *rules* (Bullinaria and Chater, 1995). Nonetheless this source of evidence may be difficult to account for by a similarity-based account, at least in the case of reading (but for an attempt see Glushko, 1979).

4.1.4.14. Over-regularization. Rule-based accounts of partially regular domains (such as the mapping between spelling and pronunciation or the English past-tense) divide knowledge into two components: a set of rules and a list of exceptions to those rules. Hence, over-extensions of rules might occur, where they should have been blocked by an exception (e.g. the past tense of go is given as goed). Such errors are observed in learning (Ervin, 1964) and have been taken as evidence for rules. Caution is necessary, in that exemplar models overgeneralize both irregular and regular past-tense forms in a manner which depends only on an item's location in phonological space (Hahn et al., 1997). Even where over-regularization does not seem to depend on close similarity to other regular items, connectionist models which are neither rule- nor similarity-based might behave appropriately (e.g. Rumelhart and McClelland, 1986; Plunkett and Marchman, 1991 and Seidenberg and McClelland, 1989; Bullinaria, 1994) although controversy on the (overall) adequacy of *these* models remains (Pinker and Prince, 1988; Forrester and Plunkett, 1994; Westermann and Goebel, 1994; Marcus et al., 1995; Nakisa and Hahn, 1996).

4.1.4.15. Summary. Data from neuropsychology and over-regularization may present problems for similarity-based models in specific contexts (e.g. reading or inflection). These criteria, however, do not appear to distinguish between rule-based and non-matching connectionist accounts. Evidence from 'memory failure' may provide a useful line of evidence, although this has currently not been explored empirically. All of these criteria depend on specific rules and instance-similarities, although in the case of reading and inflection these can be based on large bodies of theoretical and empirical work.

4.1.5. The strength of experimental evidence

We have seen that most experimental criteria distinguish specific, fully elaborated theories in particular domains more effectively than they decide between the entire classes of rule- and similarity-based models. In distinguishing different classes, although individually decisive empirical tests are difficult to provide, *convergence* of several criteria may be persuasive (Smith et al., 1992). Thus, testing between the classes of rule- and similarity-based processes seems possible, but may require integration of a range of sources of data.

4.2. Computational criteria

We have so far focused on experimental evidence and ignored the computational issues concerning the relative merits of rule- and similarity-based processing. Computational considerations are crucial, however, because any viable cognitive theory must be computationally viable.

Moreover, computational constraints provide a general perspective on rule and similarity which contrasts usefully with that from the experimental literature. In experiments such as those considered above, there is typically an inverse relation between experimental precision and generality of the result. Therefore, the construc-

tion of ever more specific experimental contexts and tasks runs the risk of contributing relatively little to our understanding of rules and similarity in normal cognition. This does not mean that experimental studies should be abandoned, but it does imply that we should pay close attention to general considerations concerning the *plausibility* of rule- and similarity-based models in normal thought. Computational constraints provide an important class of such general considerations.

Specifically, the debate between rule- and similarity-based processes in cognitive science can draw on the insights and generalizations derived from experience in AI and machine learning of attempting to use each approach in practical contexts. The lessons from computation have been little recognized in psychology. However, we suggest, these lessons provide a vital complementary source of evidence in evaluating the plausibility of rule- and similarity-based accounts of human cognition.

4.2.1. Are theories possible?

Where theories – collections of rules – are available, they can be remarkably effective, as evidenced by spectacular predictive and explanatory successes in many areas of science. The central challenge for a rule-based approach, however, is actually determining rules which adequately capture our common-sense knowledge of the world. The problem has proved to be very hard. It has required enormous intellectual effort even to provide adequate axioms for set theory and arithmetic[12]. Very few aspects of scientific knowledge have been more than partially formalized, and constructing theories for common-sense knowledge appears still more difficult. One problem is that common-sense knowledge does not appear to break up into separate domains. Thus, trying to provide rules for parts of knowledge seems to lead inevitably to the endless task of capturing the whole of human knowledge. This is what Fodor calls the *isotropy* of common-sense knowledge (Fodor, 1983). Another problem is that common-sense rules almost always have exceptions (Reiter, 1980), which raises enormous technical and conceptual difficulties (McDermott, 1987; Oaksford and Chater, 1991, 1993). Moreover, even given a set of rules, there remains the problem of how these rules are applied to specific instances. This too appears to depend on vast amounts of background knowledge, in ways that are not at all well-understood (Oaksford and Chater, 1991; Pickering and Chater, 1995).

For these reasons, AI has not been able to provide a feasibility proof of the claim that knowledge is represented in terms of rules by building general purpose rule-based systems. Although ‘expert-systems’ based on rules have been developed for highly specialized domains, e.g. DENDRAL (Feigenbaum, 1977), MYCIN (Shortliffe, 1976) or ASSESS (Dayal et al., 1993), ‘scaling-up’ to real world materials has not been achieved. Thus, it seems that a ‘pure’ rule-based approach to cognition is unlikely to be viable.

[12]Frege’s formalization of set theory, which appeared directly to reflect basic intuitions turned out to be inconsistent; and rnoreover Gödel showed that a complete, consistent axiomatization of arithmetic is impossible (Boolos and Jeffrey, 1988). Both results suggest that formalization of human knowledge may also encounter unexpected difficulties in other domains.

4.2.2. The power of similarity-based reasoning

Reasoning by similarity to past instances can be applied even in domains which are little understood and where no theory is available, where theories are partial or there are competing theories (as in law) (Ashley, 1990; Porter et al., 1990). Additionally, re-using entire past 'problem-solutions' in domains such as problem-solving or planning, can lead to faster processing than continually reasoning from first principles (Schank, 1982; Kolodner, 1991, 1992). The ability to cope with 'partial' theories is of central psychological importance, because complete common-sense theories may not be feasible, as outlined above. Indeed, even currently popular 'theory-based views' in psychology, which emphasise the role of general knowledge in categorization, claim only that *partial theories* are brought to bear, thus leaving an important role for similarity-based reasoning (Hahn and Chater, 1997).

Furthermore, similarity-based reasoning can be highly effective. For example, the simplest such algorithm, nearest neighbor, has excellent asymptotic classification accuracy in comparison with other inference methods (Cover and Hart, 1967).[13] The prospect of general and effective reasoning has fuelled enormous interest in similarity-based reasoning in AI.

In practice, however, the situation is not quite so ideal. For example, nearest neighbor methods typically require vast numbers of past instances to achieve good performance on complex problems. Moreover, similarity-based methods are dramatically impaired by redundant or irrelevant features, which have as much effect on similarity computation as the crucial ones, causing inaccurate performance (Wettschereck and Aha, 1995) and slow learning (Langley and Sage, 1994).[14] This has prompted research into so-called 'knowledge-poor' feature-selection algorithms (Aha and Bankert, 1994; Wettschereck and Aha, 1995), to choose or preferentially weight relevant features. No such method can learn optimal weight settings for all tasks, however, (Mitchell, 1990; Wettschereck et al., 1995) and, where there are only few past instances, invoking background knowledge is the only option. Thus, sophisticated CBR systems in AI (Branting, 1989; Ashley, 1990) rely on massive, knowledge-based preprocessing. The situation gets worse, where past cases are so sparse that their solutions require significant *adaptation*, such as in case-based planning (Kolodner, 1991, 1992) – a step which also introduces other forms of inference. Thus, 'scaled-up' similarity-based accounts must be supplemented with accounts of how background knowledge affects feature-weighting, which is currently poorly understood (Hahn and Chater, 1996). It seems certain that such an account will also have to integrate other, non-similarity-based, forms of inference, and that a 'pure' similarity-based account of cognition, like a pure 'rule-based' account is not viable.

[13]Asymptotically, the single nearest neighbor algorithm has (assuming smoothness) a probability of error which is less than twice the Bayes probability of error and thus less than twice the probability of error of any *other* decision rule, non-parametric or otherwise, based on the infinite sample set (Cover and Hart, 1967).

[14]The number of instances needed for nearest neighbor to reach a given level of accuracy grows exponentially with the number of irrelevant features (Langley and Sage, 1994).

4.2.3. Knowledge revision

People can clearly learn both from *experience* and from *being told* about the world. These two sources of information fit very differently with the two classes of model.

Rule-based models have the potential of adding new general information directly into the rule-base (i.e. as if the system has been *told* new information) and can interact productively with existing rules. However, this very generality means that learning from *experience* is very difficult, because there are typically many alternative changes to the rule-base that can capture new 'data.' This is analogous to the problem of theory induction or revision in the light of new data, which has notoriously resisted formal treatment. This problem has been profoundly problematic in AI; we already noted that 'expert systems' have been developed successfully only within very restricted domains, but even here, they are not the result of automated learning procedures. Rather, they are based on 'knowledge engineering,' i.e. human compilation of relevant domain knowledge into a computer accessible form. Only very limited versions of the problem of rule-induction have been addressed in machine learning, such as the induction of simple logical conjunctions from instances described as simple property lists (Langley, 1996) although more complex structures have increasingly been studied (Muggleton, 1992). Finally, rule-based systems face the problem of dealing with inconsistency between rules (for instance, as a result of 'noisy data' or of exceptions) which can potentially lead to complete inferential anarchy[15].

Similarity-based models, by contrast, can learn from experience simply by adding new instances to the data base. However, learning from being told information which covers large areas of the domain at a stroke, which is so effective for rule-based systems, is not possible for an instance-based system. On the other hand, by reasoning 'locally' from nearby instances, rather than from the global predictions of an entire system of rules, instance-based systems neither need to be globally revised, nor do they run the risk of logical inconsistency and the resulting inferential chaos. The complementary strengths of both types of system suggest integration of both.

5. Conclusions: integrating rules and similarity

In this paper, we have explicated the core distinction between rule- and similarity-based generalization, based on how representations of novel items are matched to stored representations. This core distinction is what is necessarily implied wherever the terms are contrasted without further specification. In doing so, we have resolved three difficulties with the initial intuitive distinction: we have shown that apparent mimicry arguments do not apply, we have provided clear criteria for deciding intuitively unclear cases and we have provided a clear target for empirical investigation. Moreover, we have provided an organizing framework which positions both

[15]In classical logic, all propositions and their negations can be derived from an inconsistent set of rules. Non-classical, 'para-consistent' logics, which seek to contain inference from contradictions, have therefore become a major topic of research (Touretzky, 1986; Smolenov, 1987).

rule- and similarity-based generalization in a way that shows the alternatives to both and allows visualisation of the effects of adopting further constraints on 'rule' or 'similarity' for the empirical problem of distinguishing between them.

We have also investigated the power of various experimental tests which have frequently been used to distinguish 'rules' or 'similarity' without further specification. This has revealed that these tests are not individually decisive, although convergent evidence from several sources may be compelling. We have also argued, however, that computational considerations drawn from AI provide valuable additional support in evaluating the plausibility of either account. We draw from AI the moral that pure rule- and similarity-based mechanisms appear not to be computationally viable for solving real-world problems and that neither viewpoint accounts for the human ability to learn both by example and from instruction. Both types of computational consideration suggest that the psychological concern with deciding between the two viewpoints may be misguided. Instead, it may be crucial to understand how the two can be integrated, combining the strengths of both.

This view is reflected in an increasing interest in hybrid systems within AI (Rissland and Skalak, 1991; Rissland et al., 1993). It also sits well with the not uncommon finding of both rule and similarity effects in recent experimental work on category learning reported in Section 4.1.1 above. Furthermore, given the difficulties of finding complete theories from which all desired instances can be deduced, it is also the most suggestive interpretation (Hahn and Chater, 1997) of experimental evidence in support of the theory-based view of conceptual structure (e.g. Medin and Wattenmaker, 1987). Finally, the need for interaction between the two processes is suggested by considering the structure of the law, next to science the most elaborate and explicit system we have developed for dealing with everyday life. The law displays both instance- and rule-based reasoning in the form of precedent and statute. While legal systems differ regarding the relative weight they place on each of these factors (e.g. the Anglo-American tradition emphasises similarity to past cases and the continental tradition emphasises rules), the 'blend' of both is common to all western legal systems.

These considerations suggest that rules and similarity both have their respective roles, not just side by side, with similarity covering some domains and rules others, or 'doubling up' in parallel (Sloman, 1996), but in an *active interplay* within a single task. The idea that rules and similarity might operate together is frequently suggested, even by advocates of mental rules (e.g. Smith et al., 1992; Marcus et al., 1995); and where real-world inference has been subjected to psychological explanation (Pennington and Hastie, 1993), a complex interplay of a variety of types of inference has been implicated. This suggests a shift of emphasis in future research, from pitting rules against similarity toward experimental and computational investigation of the potential interplay of rules and similarity in cognition.

Acknowledgements

The authors would like to thank Jacques Mehler, Steven Sloman and three anon-

ymous reviewers for their detailed and valuable comments on an earlier version of this manuscript, and Andreas Schöter, Andrew Gillies and Martin Redington for helpful discussion. Ulrike Hahn was funded by ESRC Grant No. R004293341442. Nick Chater was partially supported by ESRC Grant No. R000236214. The research reported in this article is based on Ulrike Hahn's doctoral dissertation and was, in part, carried out while the authors were at the Department of Experimental Psychology, University of Oxford.

References

Aamodt, A., Plaza, E., 1994. Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Communications 7, 39–59.

Aha, D., 1997. Editorial for the special issue: lazy learning. Artificial Intelligence Review 11, 7–10.

Aha, D., Bankert, R., 1994. Feature selection for case-based classification of cloud types: an empirical comparison. In: Proceedings of the AAAI-94 Workshop on Case-Based Reasoning.

Aha, D., Kibler, D., Albert, M., 1991. Instance-based learning algorithms. Machine Learning 6, 37–66.

Allen, S., Brooks, L., 1991. Specializing the operation of an explicit rule. Journal of Experimental Psychology: General 120, 3–19.

Altmann, G., Dienes, Z., Goode, A., 1995. On the modality independence of implicitly learned grammatical knowledge. Journal of Experimental Psychology: Learning, Memory and Cognition 21, 899–912.

Anderson, J., 1983. The Architecture of Cognition. Harvard University press, Cambridge, MA.

Andrews, R., Diederich, J., Tickle, A., 1995. A survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems 8, 373–389.

Ashley, K., 1990. Modeling Legal Argument – Reasoning with Cases and Hypotheticals. MIT Press, Cambridge, MA.

Barsalou, L., 1990. On the indistinguishability of exemplar memory and abstraction in category representation. In: Srull, T.K., Wyer, R.S. (Eds.), Advances in Social Cognition, Vol. III, Content and Process Specifity in the Effects of Prior Experiences. Erlbaum, Hillsdale, NJ, pp. 61–88.

Bates, E., Elman, J., 1993. Connectionism and the study of change. In: Johnson, M. (Ed.), Brain Development and Cognition. Blackwell, Oxford.

Berry, D., Broadbent, D., 1984. On the relationship between task performance and associated verbalizable knowledge. The Quarterly Journal of Experimental Psychology 86a, 209–231.

Berry, D., Broadbent, D., 1988. Interactive tasks and the implicit-explicit distinction. British Journal of Psychology 79, 251–271.

Boolos, G., Jeffrey, R., 1988. Computability and Logic, third ed. Cambridge University Press, Cambridge.

Braine, M., 1978. On the relation between the natural logic of reasoning and standard logic. Psychological Review 85, 1–21.

Branting, K., 1989. Integrating generalizations with exemplar-based reasoning. In: Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society Ann Arbor, Michigan. Erlbaum, Hillsdale, NJ, pp. 139–146.

Branting, K., 1991. Integrating Rules and Precedents for Classification and Explanation. Ph.D. thesis, University of Texas at Austin.

Brooks, L., Vokey, J., 1991. Abstract analogies and abstracted grammars: comments on Reber (1989) and Mathews et al. (1989). Journal of Experimental Psychology: General 120, 316–323.

Brooks, R., 1991. Intelligence without representation. Artificial Intelligence 47, 139–159.

Brown, M., Filer, N., 1995. Beauty vs. the beast: the case against massively parallel retrieval. In: First United Kingdom Case-Based Reasoning Workshop. Springer Verlag, in press.

Bullinaria, J., 1994. Internal representations of a connectionist model of reading aloud. In: Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society. Erlbaum, Hillsdale, NJ.

Bullinaria, J., Chater, N., 1995. Connectionist modelling: implications for cognitive neuropsychology. Language and Cognitive Processes 10, 227–264.

Chater, N., Oaksford, M., 1996. The falsity of folk theories: implications for psychology and philosophy. In: O'Donohue, W., Kitchener, R. (Eds.), Psychology and Philosophy: Interdisciplinary Problems and Responses. Sage, London.

Cheng, P., Holyoak, K., 1985. Pragmatic reasoning schemas. Cognitive Psychology 17, 293–328.

Chomsky, N., 1980. Rules and representations. The Behavioral and Brain Sciences 3, 1–61.

Chomsky, N., 1986. Knowledge of Language: Its Nature, Origin, and Use. Prager, Westport, CT.

Collins, A., 1992. On the paradox Kripke finds in Wittgenstein. Midwest Studies in Philosophy XVII, 74–88.

Cosmides, L., 1989. The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. Cognition 31, 187–276.

Cost, S., Salzberg, S., 1993. A weighted nearest neighbour algorithm for learning with symbolic features. Machine Learning 10, 57–78.

Cover, T., Hart, P., 1967. Nearest neighbour pattern classification. IEEE Transactions on Information Theory 13, 21–27.

Davies, M., 1995. Two notions of implicit rule. In: Tomberlin, J. (Ed.), Philosophical Perspectives, Vol. 9, AI, Connectionism, and Philosophical Psychology. Ridgeview, Atascadero, CA.

Dayal, S., Harmer, M., Johnson, P., Mead, D., 1993. Beyond knowledge representation: commercial uses for legal knowledge bases. In: Proceedings of the Fourth International Conference on Artificial Intelligence and Law. ACM, New York, NY.

Delosh, E., 1993. Interpolation and extrapolation in a functional learning paradigm. Purdue Mathematical Psychology Program, Purdue University.

Dreyfus, H., 1992. What computers still can't do - a critique of Artificial Reason (third ed.). MIT Press, Cambridge, MA.

Dulaney, D., Carlson, R., Dewey, G., 1984. A case of syntactical learning and judgement: how conscious and how abstract? Journal of Experimental Psychology: General 113, 541–555.

Ervin, S., 1964. Imitation and structural change in children's language. In: Lenneberg, E. (Ed.), New Directions in the Study of Language. MIT Press, Cambridge, MA.

Feigenbaum, E., 1977. The art of Artificial Intelligence: themes and case studies of knowledge engineering. In: Proceedings of IJCAI-77.

Fodor, J., 1983. Modularity of Mind. Bradford Books, London, UK; MIT Press, Cambridge, MA.

Forrester, N., Plunkett, K., 1994. The inflectional morphology of the Arabic broken plural: a connectionist account. In: Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society. Erlbaum, Hillsdale, NJ.

Funnell, E., 1983. Phonological processing in reading: new evidence from acquired dyslexia. British Journal of Psychology 74, 159–180.

Gentner, D., 1983. Structure-mapping: a theoretical framework for analogy. Cognitive Science 7, 155–170.

Gentner, D., 1989. The mechanisms of analogical learning. In: Vosniadou, S., Ortony, A. (Eds.), Similarity and Analogical Reasoning. Cambridge University Press, Cambridge, UK.

Gentner, D., Forbus, K.D., 1991. MAC/FAC: a model of similarity-based retrieval. In: Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society. Erlbaum, Hillsdale, NJ, pp. 504–509.

Gentner, D., Markman, A., 1994. Structural alignment in comparison: no difference without similarity. Psychological Science 5, 152–158.

Ginet, C., 1992. The dispositionalist solution to Wittgenstein's problem about understanding a rule: answering Kripke's objections. Midwest Studies in Philosophy XVII, 53–88.

Ginsberg, M., 1987. Readings in Nonmonotonic Reasoning. Morgan Kaufmann, San Mateo, CA.

Glushko, R., 1979. The organization and activation of orthographic knowledge in reading aloud. Journal of Experimental Psychology: Human Performance and Perception 5, 674–691.

Goldstone, R., 1994a. The role of similarity in categorization: providing a groundwork. Cognition 52, 125–157.

Goldstone, R., 1994b. Similarity, interactive activation, and mapping. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 3–28.
Goodman, N. (1972). Seven Strictures on Similarity. In: Problems and Projects. Bobbs Merill, Indeanapolis.
Goswami, U., Bryant, P., 1990. Phonological Skills and Learning to Read. Erlbaum, Hillsdale, NJ.
Hadley, R., 1993. The 'explicit/implicit' distinction. Technical report CSS-IS TR93–02. Simon Frasier University, Burnaby BC, Canada.
Hahn, U., 1996. Cases and Rules in Categorization. Ph.D. thesis, University of Oxford, UK.
Hahn, U., Chater, N., 1996. Understanding similarity: a joint project for psychology, case-based reasoning, and law. Artificial Intelligence Review, in press.
Hahn, U., Chater, N., 1997. Concepts and similarity. In: Lamberts, K., Shanks, D. (Eds.), Knowledge, Concepts, and Categories. Psychology Press/MIT Press, Hove, UK.
Hahn, U., Nakisa, R., Plunkett, K., 1997. The dual-route model of the English past-tense: another case where defaults don't help. In: Proceedings of the GALA '97 Conference on Language Acquisition.
Haugeland, J., 1985. Artificial Intelligence: The Very Idea. MIT Press, Cambridge, MA.
Hayes, P., 1979. The naive physics manifesto. In: Michie, D. (Ed.), Expert Systems in the Micro-electronic Age. Edinburgh University Press, Edinburgh.
Herbig, B., Wess, S., 1992. Ähnlichkeit und Ähnlichkeitsmasse. In: *Fall-basiertes Sehliessen – Eine Übersicht*. SEKI Working papers SWP-92–08,, University of Kaiserslautern, Germany.
Herrnstein, R., 1990. Levels of stimulus control: a functional approach. Cognition 37, 133–166.
Inhelder, B., Piaget, J., 1958. The Growth of Logical Reasoning. Basic Books, New York.
Johnson-Laird, P., Byrne, R., 1991. Deduction. Lawrence Erlbaum, Hillsdale, NJ.
Jordan, M., 1986. An introduction to linear algebra and parallel distributed processing. In: Rumelhart, D., McClelland, J. (Eds.), Parallel Distributed Processing:explorations in the Microstructure of Cognition, Vol 1: Foundations. MIT press, Cambridge, MA.
Koh, K., Meyer, D., 1991. Function learning: induction of continuous stimulus-response relations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, 811–836.
Kolodner, J., 1991. Improving human decision making through case-based decision aiding. AI Magazine, 52–68.
Kolodner, J., 1992. An introduction to case-based reasoning. Artificial Intelligence Review 6, 3–34.
Komatsu, L., 1992. Recent views of conceptual structure. Psychological Bulletin 112, 500–526.
Kripke, S.A., 1982. Wittgenstein on Rules and Private Language. Blackwell, Oxford.
Kruschke, J., 1992. ALCOVE: an exemplar-based connectionist model of category learning. Psychological Review 99, 22–44.
Lamberts, K., 1995. Categorization under time pressure. Journal of Experimental Psychology: General 124, 161–180.
Langley, P., 1996. Elements of Machine Learning. Morgan Kaufmann, San Francisco, CA.
Langley, P., Sage, S., 1994. Oblivious decision trees and abstract cases. In: Working Notes of the AAAI-94 Workshop on Case-Based Reasoning. AAAI Press.
Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., Woest, A., Pinker, S., 1995. German inflection: the exception that proves the rule. Cognitive Psychology 29, 189–256.
McCarthy, R., Warrington, E., 1986. Phonological reading: phenomena and paradoxes. Cortex 22, 868–884.
McDermott, D., 1987. A critique of pure reason. Computational Intelligence 3, 151–160.
McDowell, J., 1984. Wittgenstein on following a rule. Synthese 58, 325–363.
Medin, D.L., Wattenmaker, W., 1987. Category cohesiveness, theories, and cognitive archaeology. In: Neisser, U. (Ed.), Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization. Cambridge University Press, Cambridge, UK.
Medin, D., Schaffer, M., 1978. Context theory of classification learning. Psychological Review 85, 207–238.
Mitchell, T., 1990. The need for biases in learning generalizations. In: Shavlik, J., Dietterich, T. (Eds.), Readings in Machine Learning. Morgan Kaufmann, San Mateo, CA.
Muggleton, S., 1992. Inductive Logic Programming. Academic Press, New York.

Myllymäki, P., Tirri, H., 1993. Massively parallel case-based reasoning with probabilistic similarity metrics. In: First European Workshop on Case-Based Reasoning. Springer Verlag, Berlin.

Nakisa, R.C., Plunkett, K., Hahn, U., 1998. A cross-linguistic comparison of single and dual-route models of inflectional morphology. In: Broeder, P., Murre, J. (Eds.), Cognitive Models of Language Acquisition. MIT Press, Cambridge, MA, in press.

Nakisa, R., Hahn, U., 1996. Where defaults don't help: the case of the German plural system. In: Proceedings of the 18th Annual Meeting of the Cognitive Science Society. Erlbaum, Mahwah, NJ, pp. 177–182.

Newell, A., 1963. The chess machine. In: Sayre, K., Crosson, F. (Eds.), The Modeling of the Mind. Notre Dame University Press, South Bend, IN.

Newell, A., 1991. Unified Theories of Cognition. Cambridge University Press, Cambridge, UK.

Newell, A., Simon, H., 1972. Human Problem Solving. Prentice-Hall, Englewood Cliffs, NJ.

Newell, A., Simon, H., 1990. Computer science as empirical enquiry: symbols and search. In: Bodenz, M. (Ed.), The Philosophy of Artificial Intelligence. Oxford University Press, Oxford, UK.

Nisbett, R., Wilson, T., 1977. Telling more than we can know: verbal reports on mental processes. Psychological Review 8, 231–259.

Nosofsky, R., 1984. Choice, similarity and the context theory of classification. Journal of Experimental Psychology: Learning, Memory and Cognition 10, 104–114.

Nosofsky, R., 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory and Cognition 14, 700–708.

Nosofsky, R., Clark, S., Shin, H., 1989. Rules and exemplars in categorization, identification, and recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition 15, 282–304.

Nosofsky, R., 1992. Exemplars, prototypes, and similarity rules. In: Healy, A., Kosslyn, S., Shiffrin, R. (Eds.), From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes. Erlbaum, Hillsdale, NJ.

Oaksford, M., Chater, N., 1993. Mental models and the tractability of everyday reasoning. Behavioural and Brain Sciences 16, 360–361.

Oaksford, M., Chater, N., 1991. Against logicist cognitive science. Mind and Language 6, 1–38.

Pavlov, I., 1927. Conditional Reflexes. Oxford University Press, London, UK.

Pennington, N., Hastie, R., 1993. Reasoning in explanation-based decision making. Cognition 49, 123–163.

Pickering, M., Chater, N., 1995. Why cognitive science is not formalized folk psychology. Minds and Machines 5, 309–337.

Pinker, S., Prince, A., 1988. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. Cognition 28, 73–193.

Plunkett, K., Marchman, V., 1991. U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. Cognition 38, 43–102.

Porter, B., Bareiss, R., Holte, R., 1990. Concept learning and heuristic classification. Artificial Intelligence 45, 229–263.

Posner, M., Keele, S., 1970. Retention of abstract ideas. Journal of Experimental Psychology 83, 304–308.

Putnam, H., 1974. The 'corroboration' of theories. In: Schilpp, P. (Ed.), The Philosophy of Karl Popper, Vol. 1. Open Court Publishing.

Pylyshyn, Z., 1984. Computation and Cognition. MIT Press, Cambridge, MA.

Quine, W., 1960. Word and Object. MIT Press, Cambridge, MA.

Reber, A.S., 1989. Implicit learning and tacit knowledge. Journal of Experimental Psychology: General 118, 219–235.

Redington, M., 1996. What is learnt in Artificial Grammar Learning? Ph.D. thesis, Department of Experimental Psychology, University of Oxford.

Redington, M., Chater, N., 1994. The guessing game: a paradigm for artificial grammar learning. In: Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society. Erlbaum, Hillsdale, NJ.

Redington, M., Chater, N., 1996. Transfer in artificial grammar learning: a re-evaluation. Journal of Experimental Psychology: General 125, 123–138.
Reed, S., 1972. Pattern recognition and categorization. Cognitive Psychology 3, 382–407.
Reiter, R., 1980. A logic for default reasoning. Artificial Intelligence 13, 81–132.
Rips, I., 1994. The Psychology of Proof. MIT Press, Cambridge, MA.
Rissland, E., Skalak, D., 1991. CABARET: rule interpretation in a hybrid architecture. International Journal of Man-Machine Studies 34, 839–887.
Rissland, E., Skalak, D., Friedman, M., 1993. BankXX: A program to generate argument through case-based search. In: Proceedings of the Fourth International Conference on Artificial Intelligence and Law, ACM, New York, NY.
Rosch, E., Mervis, C., Gray, W., Johnson, D., Boyes-Braem, P., 1976. Basic objects in natural categories. Cognitive Psychology 8, 382–439.
Ross, B., 1984. Remindings and their effects in learning a cognitive skill. Cognitive Psychology 16, 371–416.
Ross, E., 1987. This is like that: the use of earlier problems and the separation of similarity effects. Journal of Experimental Psychology: Learning, Memory and Cognition 13, 629–637.
Ross, B., Kennedy, P., 1990. Generalizing from the use of earlier exemplars in problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 42–55.
Rumelhart, D., McClelland, J., 1986. On learning past tenses of English verbs. In: Rumelhart, D., McClelland, J. (Eds.), Parallel Distributed Processing, Vol 2: Psychological and Biological Models. MIT press, Cambridge, MA.
Rumelhart, D., Todd, P., 1993. Learning and connectionist representations. Attention and Performance, pp. 3–30.
Rumelhart, D., Zipser, D., 1985. Feature discovery by competitive learning. Cognitive Science 9, 75–112.
Schank, R., 1982. Dynamic Memory: A Theory of Learning in Computers and People. Cambridge University Press, Cambridge, UK.
Searle, J., 1980. Rules and causation. Behavioral and Brain Sciences 3, 1–61.
Seidenberg, M., McClelland, J., 1989. A distributed, developmental model of word recognition and naming. Psychological Review 96, 523–568.
Seifert, C., 1989. Analogy and case-based reasoning. In: Proceedings: Case-Based Reasoning Workshop. Morgan Kaufmann, San Mateo, CA.
Selfridge, O., 1959. Pandemonium: a paradigm for learning. In: Office, L.H.S. (Ed.), Symposium on the Mechanization of Thought Processes.
Shallice, T., 1988. From Neuropsychology to Mental Structure. Cambridge University Press, Cambridge, UK.
Shanks, D., 1995. Rule induction. In: The Psychology of Associative Learning. Cambridge University Press, Cambridge.
Shanks, D., John, M.S., 1994. Characteristics of dissociable human learning systems. Behavioral and Brain Sciences 17, 367–395.
Shepard, R., 1957. Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. Psychometrika 22, 325–345.
Shepard, R., 1980. Multidimensional scaling, tree-fitting, and clustering. Science 210, 390–399.
Shieber, S.M., 1986. An Introduction to Unification-Based Approaches to Grammar. Center for the Study of Language and Information, Stanford, CA.
Shortliffe, E., 1976. Computer-based Medical Consultations: MYCIN. Elsevier, New York.
Sidman, M., Tailby, W., 1982. Conditional discrimination vs. matching to a sample: an expansion of the testing paradigm. Journal of the Experimental Analysis of Behavior 37, 5–22.
Sloman, S., 1996. The empirical case for two systems of reasoning. Psychological Bulletin 119, 3–22.
Smith, E., Langston, C., Nisbett, R., 1992. The case for rules in reasoning. Cognitive Science 16, 1–40.
Smith, E., Medin, D., 1981. Categories and Concepts. Harvard University Press, Cambridge, MA.
Smolenov, H., 1987. Paraconsistency, paracompleteness and intentional contradictions. Journal of Non-classical Logic 4, 5–36.
Touretzky, D., 1986. The Mathematics of Inheritance Systems. Morgan Kaufman, Los Altos, CA.

Touretzky, D., Hinton, G., 1988. A distributed connectionist production system. Cognitive Science 12, 423–466.

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

Vaughan, W., 1988. Formation of equivalence sets in pigeons. Journal of Experimental Psychology: Animal Behavior Processes 14, 36–42.

Vokey, J., Brooks, L., 1992. Salience of item knowledge in learning artificial grammars. Journal of Experimental Psychology: Learning, Memory, and Cognition 18, 328–344.

Vokey, J., Brooks, L., 1994. Fragmentary knowledge and the processing specific control of structural sensitivity. Journal of Experimental Psychology: Learning, Memory and Cognition 18, 1504–1510.

Wason, P., 1968. Reasoning about a rule. Quarterly Journal of Experimental Psychology 20, 273–281.

Westermann, G., Goebel, R., 1994. Connectionist rules of language. In: Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Society. Erlbaum, Hillsdale, NJ, pp. 236–241.

Wettschereck, D., Aha, D., 1995. Weighting features. In Proceedings of the First International Conference on Case-Based Reasoning.

Wettschereck, D., Aha, D., Mohri, T., 1995. A review and comparative evaluation of feature weighting methods for lazy learning algorithms. Technical report AIC-95–012, Navy Center for Applied Research in AI, Washington DC.

Young, R., O'Shea, T., 1981. Errors in children's subtraction. Cognitive Science 5, 153–177.