Stuart Russell

# ANALOGY BY SIMILARITY

In this paper I discuss the relative merits of the logical and similarity-based approaches to reasoning by analogy. Although recent work by Davies and the author has shown that, given appropriate background knowledge, analogy can be viewed as a logical inference process, I reach the conclusion that pure similarity can provide a probabilistic basis for inference, and that, under certain assumptions concerning the nature of representation, a quantitative theory can be developed for the probability that an analogy is correct as a function of the degree of similarity observed. This theory also accords with psychological data (Shepard), and together with the logical approach promises to form the basis for a general implementation of analogical reasoning.

## 1. THE LOGICAL APPROACH

Analogical reasoning is usually defined as the argument from known similarities between two things to the existence of further similarities. Formally, I define it as any inference following the schema

$$P(S, A),\ P(T, A),\ Q(S, B) \stackrel{\text{anal}}{\longrightarrow} Q(T, B)$$

where $T$ is the *target*, about which we wish to know some fact $Q$ (the *query*); $S$ is the *source*, the analogue from which we will obtain the information to satisfy $Q$ by analogy; $P$ represents the known similarities given by the shared attribute values $A$. $P$ and $Q$ can be arbitrary predicate calculus formulae, and $A$ and $B$ stand for arbitrary tuples of objects.

An innumerable number of inferences have this form but are plainly silly; in other words, the form does not distinguish between good and bad analogical inferences. For example, both today and yesterday occurred in this week (the known similarity), yet we do not infer the further similarity that today, like yesterday, is a Friday. The traditional approach to deciding if an analogy is reasonable, apparently starting with Mill (1843), has been to say that each similarity observed contributes some extra evidence to the conclusion; this leads naturally to the assumption that the most suitable source analogue is the one which has the greatest similarity to the target; presumably, one can take into account differences in the same way. Thus similarity becomes a measure on the *descriptions* of the source and target. However one defines the similarity measure, it is trivially easy to produce counterexamples to this assumption. Moreover, Tversky's studies (1977) show that similarity does not seem to be the simple, two-argument function this naïve theory assumes. One can convince oneself of this by trying to decide which day is most similar to today.

In the philosophical literature on analogy, several authors have noted the inadequacy of 'similarity-counting' arguments as the basis for analogy, particularly since many analogies are extremely convincing. One approach to logical justification proposes that knowledge of the rule $\forall \underline{x}[P(\underline{x}, A) \Rightarrow Q(\underline{x}, B)]$ is needed for an analogy to be sound, but such knowledge would render the analogue $S$ logically superfluous. Keynes (1957), Uemov (1964), Anderson (1969) and Nagel (1961) all pointed out this possibility for justified analogy, and all stated that no other possibility existed. The 'trivial' nature of such analogies may have led Greiner (1985) to *define* analogy as necessarily non-logical. Hesse (1966) noted the importance of *relevance* of the known similarities to the inferred similarities. The theory of *determinations* (Davies, 1985; Russell, 1986c; Davies & Russell, 1987; Davies, this volume) gives a first-order definition to the notion of relevance.* Given that the known similarities are (partially) relevant to the inferred similarities, the analogical inference is guaranteed to be (partially) justified. The fact that P is relevant to Q is encoded as a determination, written as $P(\underline{x}, \underline{y}) \succ Q(\underline{x}, \underline{z})$ and defined as

$$\forall \underline{w}\underline{x}\underline{y}\underline{z} P(\underline{w}, \underline{y}) \wedge P(\underline{x}, \underline{y}) \wedge Q(\underline{w}, \underline{z}) \Rightarrow Q(\underline{x}, \underline{z}).$$

When the reasoner has this kind of background information available, attention can be directed to those similarities that are relevant to the problem at hand, and the justification of the conclusion is logical in nature; the overall degree of similarity no longer plays a part in the process.

I am thus proposing that at least one aspect of a successful analogical reasoning system consists of a knowledge-based, deductive process (or, in the case of partial determinations, a probabilistic process). Determinations seem to be a common and useful form of knowledge, and we can ascribe to determinations the same epistemological status and heuristic utility as we do to the typical universally-quantified rules in a rule-based expert system. It would be interesting to perform psychological experiments to ascertain subjects' knowledge of determinations, and to design knowledge engineering methods for eliciting them from experts. In (Russell, 1986c) I give methods for inductive acquisition of determinations and for their use in a logical inference system. The crucial argument for the value of determination-based analogy is that determinations represent that class of regularities whose extrapolation takes the form of analogical reasoning; without the ability to detect and use determinations, a system is simply impoverished in its inferential power. The question remains as to whether other forms of analogy have a rational justification, particularly in the light of the common conception of analogy as only a plausible inference process, or as a learning method. The phrase 'learning by analogy' appears repeatedly — in fact, the study of analogy is almost universally classified as a subfield of machine learning in conferences and textbooks. In the next section we examine how this

* Goodman (1955) also identified this class of formulae in his work on induction, calling them 'overhypotheses'.

widespread belief can be reconciled with our theory.

## 2. LEARNING AND ANALOGY

Analogical inference using determinations does not constitute learning in the strict sense of acquisition of new knowledge, whether the determinations are deductive or probabilistic. There is no 'learning at the knowledge level' (Dietterich, 1986) occurring when an analogical conclusion is reached in this way; the perception of analogy as learning may simply have arisen because the determination premise is not immediately obvious to introspection. The idea behind the phrase 'learning by analogy' is that similarity, in and of itself, should be enough to suggest new information that may be usefully conjectured. The 'creative' nature of analogy is often stressed. Yet no one would deny that however creative or interesting a conjecture may be, the only way we can decide whether or not to make that conjecture is to have some idea of how likely it is to be true. For otherwise, we might just as well select hypotheses at random from the space of all expressible conjectures. Words such as 'plausible' and 'conjectural' often seem to be ways of putting off the realization that ultimately we are just talking about probabilistic inference, whether the probabilities be high or low. Under this 'hard-nosed' view, we have *separated* learning and inference. We can instead take the 'soft-nosed' position, which is perhaps preferable, and say that all inference to unobserved conclusions in empirical domains is necessarily probabilistic, just as inductive generalization is probabilistic. Then the distinction might be made between inferences that extrapolate regularities to new cases and those, which we might call 'learning', that postulate new regularities or generate new beliefs by some means other than extrapolation. Analogy by similarity has been the candidate for this last possibility. In the same sense that Goodman says that our best inductive practice is a good enough justification for an inductive inference, it is possible that a refined procedure for analogy by similarity may form a primitive constituent of our inferential apparatus, in need of no further justification. However, until this step is shown to be necessary, as in the case of induction, it seems preferable not to take it.

Thus, in strict terms, the phrase 'learning by analogy' may be somewhat misleading, if analogical inference is just the extrapolation of a previously detected regularity (the determination). Again, it is possible that the analogy process may use unfounded, syntactic heuristics to produce its conclusions. The only syntactic inference rules we are allowed to use willy-nilly are those based ultimately on the semantics of the representation language, i.e., the rules relating syntax to truth. For example, Modus Ponens is based on the Tarskian semantics for predicate calculus. Syntactic rules of the type exemplified by the 'analogy by similarity' heuristic appear to be justifiable only empirically, by showing that they tend to work. Even then, one is left with the (in some cases insurmountable) problem of showing that the results are not influenced by some special features in the form or content of the knowledge base. We will now see how this might work, in a couple of different ways.

## 3. REPRESENTATIONAL JUSTIFICATION OF SIMILARITY HEURISTICS

In this section we give the first intimations, in a very simplistic fashion, of one possible direction that might be explored as a way of justifying a form of analogy by similarity.

Recall that the commonality between two 'objects' may be expressed by giving a common formula $P$ holding for both. According to the traditional view, the 'size', measured in some way, of this common formula is the basis for analogical transfer. Suppose we define a meta-linguistic predicate *Large*, indicating that its argument is, in this sense, a large formula. Then, very loosely, analogy by similarity corresponds to the axiom

$$\forall P, Q[Large(P) \Longrightarrow [P(\underline{x}, \underline{y}) \succ Q(\underline{x}, \underline{z})]].$$

Such heuristics could exist at the top of a hierarchy of determinations, to be used when no more specific knowledge is available. The *justification* for the use of such heuristics can rest on their empirical success. However, it is not hard to imagine knowledge bases and representations for which the heuristic fails miserably. Because the heuristic works only at the syntactic level, we can always construct consistent knowledge bases such that the use of the heuristic is actually deleterious. In other words, the use of the heuristic contains an implicit restriction on the possible conditions obtaining in the universe of discourse.

To remain coherent, such syntactic theories should include the representational and epistemological assumptions that allow them to work correctly, and motivate those assumptions. Such assumptions might be, for example, that only facts about certain types of object will be included, or that only certain relations are explicitly stated, or that inferential goals will tend to be of a certain type. These assumptions can be justified using a theory describing that part of the system responsible for acquiring the vocabulary and content of the knowledge base, and its relationship to the world.

Humans (and computers) could in fact possess a general-purpose similarity heuristic (possibly dependent on the things compared) which works well for the 'average' query. Let us consider an example of such a representational assumption. Psychological attunement theories regarding the way in which representations, as well as their contents, evolve to reflect underlying regularities in the environment may be one source of such heuristics. Thus, humans do in fact seem to record only the 'important' features of their experiences; what has come to be important must depend on the use to which experiences are normally put, and evolutionarily speaking those uses have been in deciding such things as edibility, dangerousness, running speed and other gross physical properties of the objects in our world. Thus, biasing similarity metrics towards simple, observable, constant, physical features is a justifiable policy for early man. Unfortunately, similar justifications have not been made for any of the similarity metrics used in AI theories of analogy. There is a large amount of work to be done before we can begin to understand fully the ways in which a system can take advantage of representational regularities in order to achieve inferential shortcuts.

## 4. A QUANTITATIVE ANALYSIS OF ANALOGY BY SIMILARITY

I now propose a second approach to the analysis of analogy by similarity, one that yields more quantitative results. We start from the case in which we are trying to solve some problem by analogy, but we know no applicable determination for the query at hand, i.e., we have no idea which of the known facts might be relevant. In this case, the theory of determinations does not apply. However, it still seems plausible that the most similar source is the best analogue; certainly, in the absence of any other information, it seems perverse to choose an analogue that is demonstrably *less* similar. What has been lacking in previous theories of analogy by similarity is any attempt to justify this assumption; the analysis in this section hopes to rectify this situation. Since an inference by analogy is still an inference, the justification must take the form of an argument as to why a conclusion from similarity is any better than a random guess; better still, the theory should be able to assign a probability to the conclusion given the truth of the premises. The object of this section is thus to compute (or at least sketch) the relationship between the measure of similarity between two objects, and the probability that they share a further, specified similarity.

The principal problems which need to be solved before such a theory can be constructed are:

1) A reasonable way must be found to circumscribe the source and target descriptions. Without this, the sets of facts to be compared are essentially without limit.
2) A similarity measure must be defined in such a way as to be (as far as possible) independent of the way in which the source and target are represented.
3) We must identify the assumptions needed to relate the similarity measure to the desired probability.

The precise similarity measure itself is not important; in fact, it is essentially meaningless. If we have a different similarity measure, we simply need to relate it in a different way to the probability of correctness of the analogy. Thus I will *not* be attempting to define a similarity measure that is more plausible than those proposed previously.

The essence of our approach is to show that analogy to a maximally similar source can be justified in the absence of any usable determination by showing that such a source is the most likely to match the target on the properties which are relevant to the query *even though the identity of these properties is unknown.* The intuition on which the analysis is based is the following: in situations where the system is extremely ignorant, there will be many determinations (causal factors) of which it is unaware. Thus some facts could be relevant to the query even if we have no direct reason to believe them so. In this case, a larger similarity serves to increase the likelihood that such factors will be taken into account, by increasing the likelihood that the relevant features will be included in the commonality.

If a source matches the target on all relevant features, an analogy from that source is assumed to be correct. For the query to be soluble at all, we require that all the features relevant to the query appear *somewhere* in the description of the target to be matched against the source. This is equivalent to saying that the formula describing the target is a sufficient determinant for the query; conversely, when a determination is known for a query its left-hand side can be used to circumscribe the facts needed in the description of the target and source for the purposes of matching. When these match completely, we have complete similarity on the relevant features and the limiting case is thus the same as the logical approach. When the match is not complete, the theory we are about to describe allows a probabilistic conclusion. Thus even a highly overconstrained determination, whose left-hand side is far too specific (i.e., contains too many features) to offer a reasonable chance of achieving the match needed for a sound analogy, is still useful for constraining the object descriptions used in similarity matching.

I first calculate the probability of a match on the relevant attributes for the simple case of an attribute-value representation where a match on any attribute is equally likely *a priori*, and I assume a fixed number of relevant features. Subsequent sections relax these assumptions to allow the theory to apply to the general case, in the process revealing the representational assumptions that underlie my analysis.

### 4.1 The simple model

A simplified model for analogy in a database is this: we have a target $T$ described by $m$ attribute-value pairs, for which we wish to find the value of another attribute $Q$. We have a number of sources $S_1 \dots S_n$ (analogues) which have values for the desired attribute $Q$ as well as for the $m$ attributes known for the target.

Define the similarity $s$ as the number of matching attribute values for a given target and source. The difference $d = m - s$. Assume that there are $r$ attributes relevant to ascertaining the value of $Q$.

Define $p(d, r)$ to be the probability that a source $S$, differing from the target on $d$ attributes, matches it on the $r$ relevant attributes. The assumption of no relevance information means that all attributes are equally likely to be relevant. We can thus calculate $p(d, r)$ using a simple combinatoric argument:

Let $N_m$ be the number of choices of which attributes are relevant such that S matches T on those attributes.

Let $N$ be the total number of choices of which attributes are relevant.

$$p(d,r) = N_m/N$$
$$= \binom{m-d}{r} \Big/ \binom{m}{r} \quad (r \geq 1)$$

For any r, this function drops off with d (= m-s), monotonically and concavely, from 1 (where d=0) to 0 (where d > m-r). Thus the most similar analogue is guaranteed to be the most suitable for analogy. Figure 1 shows $p(d,r)$ for values of $r$ of 1, 3, 5, 10, 20 with the total number of attributes $m = 30$. As we would expect, the curve narrows as $r$ increases, meaning that a higher number of relevant attributes necessitates a closer overall match to ensure that the relevant similarities are indeed present.

**Fig. 1** $p(d,r)$ for $r$ = 1,3,5,10,20.

4.2 Allowing r to vary

The assumption of a fixed value for the number of relevant features seems rather unrealistic. The most general assumption we can make is that $r$ follows a probability distribution $q_Q(r)$ which depends on the type of the query

$Q$. Thus, for example we could assume that there are equally likely to be any number of relevant features, or that three or four seems reasonable whilst 25 is unlikely. Although this introduces an extra degree of freedom into the theory, we find that the results are almost independent of what we assume about $q$. We calculate the probability of successful analogy now as a function of the source-target difference $d$ only:

$$p(d) = \sum_{r=0}^{m} q(r)p(d, r)$$

using the above formula for $p(d, r)$. For any reasonable assumption about the shape of q(r), the variation of p(d) with d remains approximately the same shape.

For $q(r) = constant$, $p(d) \sim 1/(d + 1)$

For $q(r) \propto e^{-r}$, $p(d) \sim e^{-d}$ for low d, larger for high d

For $q(r) \propto re^{-r}$, $p(d) \sim e^{-d}$ except at large d

For $q(r) = Normal(\mu = 4, \sigma = 2)$, $p(d) \sim e^{-d}$

$q(r) = constant$ $q(r) \propto e^{-r}$ $q(r) \propto re^{-r}$ $q(r) = N(4, 2)$

**Fig. 2** $p(d)$ given various assumptions about $q(r)$.

In figure 2 we show values of p(d) (plotted as dots) computed using these four assumptions of $q(r)$, with a simple exponential decay ($p(d) \propto e^{-d}$, solid line) superimposed.

4.3 Generalizing the model

We can make the simple model analyzed above applicable to any analogical task simply by allowing the 'attributes' and 'values' to be arbitrary predicate calculus formulae and terms. The assumption that a match on any of these new 'attributes' is equally likely, *a priori*, is no longer tenable, however. In this section we will discuss some ways in which the similarity measure might be modified in order to allow this assumption to be relaxed. The idea is to reduce each attribute to a collection of uniform mini-attributes; if the original assumptions hold for the mini-attributes, our problem will be solved. Unfortunately, the task is non-trivial.

The first difficulty is that we can only assume equal relevance likelihood if the *a priori* probabilities of a match on each attribute value are equal; in general, this will not be the case. In the terms of Carnap (1971), the *widths* of the regions of possibility space represented by each attribute are no longer equal. Accordingly, the simple notion of similarity as the number of matching attributes needs to be revised. If the cardinality of the range of possible values for the $i^{th}$ attribute is $k_i$, then the probability $p_i$ of a match (assuming uniform distribution) is $1/k_i$. Although $k$ will vary, we can overcome this by reducing each attribute to $\log_2 k$ mini-attributes, for which the probability of a match will be uniformly 0.5. If the original distribution is not uniform (for example, a match on the NoOfLegs attribute with value 2 is much more likely than a match with value 1), a similar argument gives the appropriate contribution as $-\log_2 p_i$ mini-attributes. This refinement may underlie the intuition that 'unusual' features are important in metaphorical transfer and analogical matching (Winston, 1978; Ortony, 1979). A generalization of this idea would deal with arbitrary probability distributions for the values of $p$, incorporating the inexact match idea of the following paragraph.

In the logical approach, the notion of one attribute value 'almost matching' another is expressed as a commonality by defining a more coarse-grained attribute, such that the two 'close' values are mapped onto the same value for the new attribute. A representation should be chosen such that determinations are expressed using the 'broadest' attributes possible, thus precise attributes are grouped into equivalence classes appropriate to the task for which we are using the similarity. In the current situation, however, we will not know what the appropriate equivalence classes are, yet we still want to take into account inexact matches on attribute values; for example, in heart disease prognosis a previous case of a 310-lb man would be a highly pertinent analogue for a new case of a 312-lb man. If the weight attribute was given accurate to 4 lbs instead of 1lb, these men would weigh the same; thus in general an inexact match on a scalar attribute corresponds to an exact match on less fine-grained scale, and the significance of the 'match' is reduced according to the log of the accuracy reduction (2 bits in this case).

A consequence of this view of the significance of an attribute leads to a constraint on the possible forms of $q(r)$: if we assume that the relevant attributes must contain at least as much *information* as the attribute $Q$ whose value they combine to predict, then we must have $q(r) = 0$ if $r$ is less than the significance value of $Q$. Here $r$, as well as the total 'attribute count' $m$ and the similarity $s$, are all measured on a scale where a one-bit attribute has a significance of 1. At first sight, it seems that we have succeeded in breaking down our complex features into uniform elements, all of which are equally likely to be relevant, so all the earlier results should still apply.

However plausible this may seem, it is simply false. The base of the logarithms chosen is of course totally arbitrary — we would still have uniform mini-attributes if we had used $\log_4$. This would mean halving our values for $m$, $r$ and $s$; but the formula for $p(d, r)$ contains combinatoric functions, so it will not scale linearly. Hence our predicted probability will depend on the base we choose for the logarithms! This is clearly unsatisfactory. What we have done is to neglect an important assumption made in using the combinatorial argument, namely that the relevant information consisted of a set of *whole features*. If we allow it to consist of a collection of sub-elements of various features, then clearly there are many more ways in which we can choose this set. The plausibility of the simple model rests in our unstated assumption that the attributes we use carve up the world in such a way as to correctly segment the various causal aspects of a situation. For example, we could represent the fact that I own a clapped-out van by saying

$$OwnsCar(SJR, 73DodgeSportsmanVanB318)$$

using one feature with a richly-structured set of values; but for most purposes a reasonable breakdown would be that I own a van (for other people's moving situations), that it's very old (for long-distance trip situations), that it can seat lots of people (for party situations), that it's a Dodge (for frequent repair situations) and that it's virtually worthless (for selling situations). Few situations would require further breakdown into still less specific features. In some sense, therefore, we will require a theory of natural kinds for features as well as for objects.

If it is the case that humans have succeeded in developing such well-tuned representations, then it is indeed reasonable for us to assume that the relevant information, which corresponds to the part of the real-world situation which is responsible for determining the queried aspect, will consist of a set of discrete features corresponding to the various possible causal factors present. This of course raises a vast throng of questions, not least of which is that of how an AI system is to ensure that its representation has the appropriate properties, or even how it can know that it does or doesn't. The subject of the semantic implications of using a particular representation is also touched upon in the concluding section of this paper.

## 5. EMPIRICAL DATA ON STIMULUS GENERALIZATION

A crucial test of whether the representational assumptions used in the above quantitative analysis are reasonable is to compare its predictions to actual human and animal performance. Psychological experiments on *stimulus generalization* are essentially measuring the subject's ability to do analogy by similarity. In these experiments, a (human or animal) subject is given an initial stimulus, to which it makes a response. If necessary, the correct response is confirmed by reinforcement. This original stimulus-response pair is the *source* in our terms. Then a second stimulus is given, which differs from the original. This represents the *target* situation, for which the subject must decide if the original response is still appropriate. The empirical probability that the subject makes the same response (*generalizes* from the original stimulus) is measured as a function of the difference between the stimuli. This probability is essentially what we are predicting from rational grounds in the above analysis.

Early results in the field failed to reveal any regularity in the results obtained. One of Shepard's crucial contributions (1958) was to realize that the similarity (or difference) between the stimuli should be measured not in a *physical* space (such as wavelength of light or pitch of sound) but in the subject's own *psychological* space, which can be elicited using the techniques of multi-dimensional scaling (Shepard, 1962). Using these techniques, Shepard obtained an approximately exponential stimulus generalization gradient for a wide variety of stimuli using both human and animal subjects. Typical results, reproduced, with kind permission, from Shepard's APA presidential address (1981), are shown in figure 3.

His own recent theory to explain these results appears in (Shepard, 1984), and has a somewhat similar flavour to that given here, although it is designed for continuous-valued stimuli. The empirical verification of the theory by Shepard's results is extremely good, in the sense that it shows that humans and animals possess a rational ability to judge similarity which has evolved or been learned, presumably, because of the optimal performance of its predictions given the available information. Shepard's explanation of the results and our own are somewhat complementary in that he deals with unanalyzed stimuli whereas our model assumes a breakdown into features. This is well-suited for our purpose of constructing a computational theory of analogy and a generally useful analogy system for AI; this is the subject of the next section.

## 6. COMBINING THE LOGICAL AND SIMILARITY-BASED APPROACHES

There seems little doubt that, given a suitable determination, determination-based analogical reasoning (DBAR) is the preferred mode of analogical reasoning, especially given the sharp fall-off in probability of correctness for the similarity-based method as the similarity decreases. We intend to further verify the similarity theory by performing analogies in an AI database of general knowledge (Lenat's CYC system; see (Lenat et al., 1986)), which will also give us an empirical form for $q(r)$. A further goal is to integrate analogy by similarity with the determination-based analogical reasoning theory to provide an analogy capablity for a general reasoning program. The integration rests on the following principles:

**Fig. 3** Plots of analogical response probability (S) against source-target difference (D), for various data, from (Shepard, 1981).

1) For either type of reasoning, we must find a determination for the given query; this may be already known, or found inductively or deductively from background knowledge.
2) If the determination is too specific to allow an exact matching source to be found, it can be used to point out broad classes of potentially relevant features; we then reason by similarity within these constraints;
3) Probabilistic determinations can add specific weights to the contributions of individual attributes to the overall similarity total;
4) Blind statistical search for new determinations is combinatorially explosive; observation of an unexpectedly high similarity can initiate a more focused search for a hitherto unknown regularity to be encoded

as a new determination.

## 7. SUMMARY

Although correct analogical reasoning requires knowledge of determinations, two other approaches show promise for the justification of analogy by similarity. The first is based on assumptions about the form and content of the system's representation of the world. Attunement in humans and animals seems to suggest that in constrained environments this approach to analogy may have promise, but it must await a better theory of representation before it can be useful. A second approach, using the idea of unidentified relevant features, seems to correspond well to the traditional idea of analogy. A quantitative relationship is developed between the degree of similarity and the probability of correctness of an analogy; the similarity measure used goes some way towards being representation-independent. When intelligent systems embodying full theories of limited rationality are built, an ability to perform analogical reasoning using both determinations and similarity will be essential in order to allow the system to use its experience profitably. Analogy by similarity also seems extremely well suited to the task of producing reliably fast, plausible answers to problems, particularly in a parallel environment.

The analysis in this paper revealed a reliance on a strong assumption about the nature of representation, namely that each attribute corresponds to an atomic 'causal factor' in the actual world. There is an echo here of the concept of *entrenchment* that Goodman uses in describing our inductive practice — only well-entrenched terms, that have frequently been involved in successful inductive hypotheses before, can be used in new inductive hypotheses. Entrenchment can be codified logically (Russell, 1986a), but a similar analysis does not yet seem possible for the representation conditions for analogy by similarity.

Entrenchment and the 'atomic causal factor' assumption are two examples of conditions on the *representation* of knowledge that can be ensured by the use of an appropriate language evolution mechanism. Given such a mechanism, inference methods that are unsound on the surface can be used reliably and efficiently, since they do not have to work with an arbitrary knowledge base. Their operation is justified by the semantics of the *presence of the terms in the language*. This is an example of what Kuhn has called *lexically-embodied knowledge*. The use of linguistic biases such as the least disjunction principle (Utgoff, 1986) in concept learning systems is another example of a syntactic inference method, but one whose logical basis has not yet been examined. A fourth, simple example is the use of the Unique Names assumption in database theory.

A first step in the process of unravelling this relationship between language and inference might be to perform a logical analysis of a given language evolution mechanism and to generate its associated syntactic inference procedure.

At present, we have very little idea how much the use of human-derived concepts in AI systems (other than pure deductive systems) contributes to their success. Consequently, we have no idea how to assure the same degree of success for an autonomous, self-evolving system. Imagining a language none of whose terms embody any knowledge is perhaps the hardest part of knowing what it is like to be a computer.

## ACKNOWLEDGMENTS

I would like to thank my advisors Doug Lenat and Mike Genesereth for fruitful discussions of these ideas, Benjamin Grosof and Devika Subramanian for helpful comments on an earlier draft, and Roger Shepard for making these ideas possible.

*Computer Science Division,*

*University of California,*

*Berkeley. 1987.*

## REFERENCES

Anderson, R. O. (1969). “Conclusive Analogical Argument”. In *Revue Internationale de Philosophie*, Vol. 23, pp. 44-57.

Carnap, Rudolf (1971). “A Basic System of Inductive Logic, Part I”. In R. Carnap and R. C. Jeffrey (Eds.) *Studies in Inductive Logic and Probability* Vol I. Berkeley, CA: University of California Press.

Davies, Todd (1985). *Analogy.* Informal Note No. IN-CSLI-85-4, Center for the Study of Language and Information, Stanford University.

Davies, Todd & Stuart Russell (1987). “A Logical Approach to Reasoning by Analogy”. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Milan, Italy.

Dietterich, Thomas G. (1986). *Learning at the Knowledge Level.* Technical Report No. 86-30-1, Computer Science Department, Oregon State University.

Greiner, Russell (1985). *Learning by Understanding Analogies.* Ph. D. thesis. Technical Report no. STAN-CS-85-1071. Stanford University.

Hesse, Mary (1966). *Models and Analogies in Science.* Notre Dame, Indiana: University of Notre Dame Press.

Keynes, John Maynard (1957). *A Treatise on Probability.* London: Macmillan.

Lenat D., Prakash, M. and Shepherd M. (1986). “CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks”. *AI Magazine* Vol. 6 No. 4.

Mill, J. S. (1843). *System of Logic* Book III Ch XX ‘Of Analogy’ in Vol. VIII of *Collected Works of John Stuart Mill.* University of Toronto Press; 1973.

Nagel, Ernest (1961). *The Structure of Science.* New York: Harcourt, Brace and World.

Ortony A. (1979). “Rôle of Similarity in Similes and Metaphors”. In Ortony A. (Ed.) *Metaphor and Thought.* Cambridge University Press, Cambridge.

Russell, Stuart J. (1986a). “Preliminary Steps toward the Automation of Induction”. In *Proceedings of the National Conference on Artificial Intelligence.* AAAI; Philadelphia, PA.

Russell, Stuart J. (1986b). “A Quantitative Analysis of Analogy by Similarity”. In *Proceedings of the National Conference on Artificial Intelligence.* AAAI; Philadelphia, PA.

Russell, Stuart J. (1986c). *Analogical and Inductive Reasoning*. Ph. D. thesis, Stanford University.

Shepard R. N. (1958). "Stimulus and Response Generalization: Deduction of the Generalization Gradient from a Trace Model". *Psychological Review* Vol. 65.

Shepard R. N. (1962). "The analysis of proximities: multidimensional scaling with an unknown distance function (Parts I and II)". *Psychometrika* Vol. 27.

Shepard, Roger (1981). APA Division 3 Presidential Address, Los Angeles, August 25.

Shepard R. N. (1984). "Similarity and a law of universal generalization". Paper presented at the annual meeting of the Psychonomic Society, San Antonio, TX.

Tversky, Amos (1977). "Features of Similarity". *Psychological Review* Vol. 84, No. 4.

Uemov A. I. (1964). "The basic forms and rules of inference by analogy". In P. Tavanec (or Tavanets) (Ed.), *Problems of the Logic of Scientific Knowledge*. Dordrecht: D. Reidel; 1970. Translated from the Russian by T. J. Blakeley; Moscow: Nauka.

Utgoff P. E. (1984). *Shift of Bias for Inductive Concept Learning*. Ph.D. thesis, Rutgers University.

Winston, Patrick H. (1978). "Learning by Creating and Justifying Transfer Frames". *Artificial Intelligence* Vol. 10, No. 4.