PRELIMINARY STEPS TOWARD THE AUTOMATION OF INDUCTION

Stuart J. Russell
Department of Computer Science
Stanford University
Stanford, CA 94305

ABSTRACT

Rational inductive behaviour is strongly influenced by existing knowledge of the world. This paper begins to elucidate the formal relationship between the base-level induction to be attempted, the direct evidence for it (positive and negative instances) and the indirect evidence (higher-level regularities in the world). By constructing a program to search the space of forms of higher-level regularity we discover some important new forms which have direct application to analogy, single-instance generalization and enumerative induction in general. We outline a theory which we hope is the first step towards the construction of powerful and robust learning systems. *

I INTRODUCTION

Ultimately, the source of all our knowledge of the world must be observation, either direct, communicated or inherited. One of the principal problems of philosophy has been to explain how this accumulation of observations can be used to fill in the gaps in our knowledge, particularly of the future. Without such an ability, rationality, which requires the prediction of the outcome of our actions, would be impossible. In AI, the problem is doubly acute: not only do we desire to understand the process for its own sake, but also without such an understanding we cannot build machines that learn. The basic answer to the problem is that we come to believe in some generally applicable rules (universals) by a process of *induction* from prior instances of their application; we then apply these rules in situations of incomplete knowledge using *deduction.* So far, so good. In AI, the two halves of the process correspond roughly to the division into the areas of machine learning and knowledge-based systems. Analogy, which seems at first sight to defy this classification, is shown in [Davies & Russell 86] to belong more to the deductive phase.

In this paper, our object is to make some progress towards a theory of induction which will prescribe, as far as is possible, the correct inductive behaviour for an intelligent system. As explained below, one essential element of this task is to explicate the way in which existing world knowledge affects a system's inductive acquisition of new knowledge. This need is pointed out in [Michalski 83]. In order to explain how present-day intelligent systems (such as ourselves) have arrived at our degree of understanding of the world, given the fact that at the beginning of evolutionary history there was no existing knowledge, our theory must provide a *formal* relationship between the system's existing knowledge and the universal to be induced; put simply, we seek a domain-independent theory. The basic problem to be solved is this: *given a mass of ground facts and no other domain knowledge, what can be inferred?*

As mentioned earlier, we perform inductions on the ground facts to obtain universals. *Enumerative induction* is just the simple process by which, from a collection of instances $a_1 \ldots a_n$ satisfying $P(a_i)$ and $Q(a_i)$, we induce the general rule $\forall x[P(x) \Rightarrow Q(x)]$. The search for a rationale for this inductive step seems to be circular: we use it because it has always worked, but the belief that this means it will work in the future requires an inductive step. This is Hume's Problem of Induction, which, according to modern interpretation, he rightly deemed to be inherently insoluble. If we could prove an enumerative induction to be valid, this would amount to prevision of the future, a scientifically dubious concept.*

Intuitively, an enumerative induction is made more certain by the discovery of further confirming instances as long as no disconfirmation occurs. This model of induction is somewhat different from the version space approach to concept learning ([Mitchell 78]), in which the generalizations produced are justified by a *linguistic bias* which limits the set of allowable generalizations so that if only one of the set is consistent with the observations then it is assumed to be true. This means that the number of confirming instances is ignored. Moreover, the factual content of the linguistic bias is neither elucidated nor motivated (but see [Utgoff 84]); in this light it is hard to view the version space approach as a form of inference. This issue is also discussed in [Dietterich 86]. The problem with which we are concerned is not just the *selection* of an appropriate generalization for some data, but the assessment of its probable truth; selection derives automatically from this if we select the most probable generalization.

In particular, we wish to investigate why one generalization may be given a great deal of credence, whilst another is regarded very suspiciously, even though they both have the same number of positive instances and no negative instances. For example, consider the case of the traveller to Italy meeting her first Italian. On hearing him speak Italian, she immediately concludes that all Italians speak Italian; yet on discovering that his name is Giuseppe, she doesn't conclude that all Italians are called Giuseppe. Clearly, the difference lies in the traveller's prior knowledge of countries, languages and names.

Goodman's classic example of grue emeralds is another case in point, which he used in [Goodman 46] to refute the early claims of the confirmation theorists (Carnap and others) that the probability of a proposition could be inferred from its instances and syntactic form alone. In Goodman's example, we

* This work was performed while the author was supported by a NATO studentship from the UK Science and Engineering Research Council, and by ONR contract N00014-81-K-0004. Computing support was provided by the SUMEX-AIM facility, under NIH grant RR-00785.

* For, despite our best predictions, the whole world could be swallowed tomorrow by a giant intergalactic toad ([Hoppe]).

are to evaluate the two inductions

1) All emeralds are green
2) All emeralds are grue

given that all the millions of emeralds observed to date are green, where *grue* means 'green if observed before tomorrow, and blue if observed thereafter'. Both have the same overwhelming collection of positive instances, but the second seems somewhat undesirable. Goodman's answer to what became known as the 'new riddle of induction' is that the rule must be *projectible*. We will return to the definition of projectibility in a later section. In spirit, his answer is the same as ours:

> "While confirmation is indeed a relation between evidence and hypotheses, this does not mean that our definition of this relation must refer to nothing other than such evidence and hypotheses. The fact is that whenever we set about determining the validity of a given projection from a given base, we have and use a good deal of other relevant knowledge." ([Goodman 83] pp. 84-5).

The object of this paper is to show what this knowledge consists of, and to show how it can be found and used to give additional confirmation to enumerative inductions. What we want is a theory which will be able to start with *any* body of knowledge of *any* world (preferably in wff form), and say which inductions are reasonable and which aren't. We therefore require that the 'other relevant knowledge' have a *syntactic* relationship to the evidence and inductive hypothesis, since otherwise the theory itself will be assuming something factual about the world, and hence will fail when applied to a world in which the factual assumption is false. In this, we strongly disagree with [Holland et al. 86], who say "In essence, our approach is to deny the sufficiency of purely syntactic accounts ... and to insist that sensible inferential rules take into account the kinds of things being reasoned about." We believe that such an approach simply begs the question of how such world-dependent rules could ultimately be acquired, except by some syntactic process; moreover, a physical system seems fundamentally incapable of performing anything but syntactic processes. Fortunately, in a formal system, logical entailment is a syntactic relationship (this is the fundamental achievement of the study of logic since Aristotle) and will play a large rôle in our theory.

If we are to build systems which observe an environment containing regularities and make use of them via the process of induction, we must be able to eliminate such spurious inductions as 'all emeralds are grue'. It might be argued that Goodman is playing the sophist here; a philosopher might wish to know why emeralds are not considered grue, but the AI pragmatist might object that this is creating difficulties for the sake of it, and that we can avoid such problems in real systems just by not coining absurd, unmotivated concepts. However, an AI system *needs* to coin new terms (see, e.g., [Lenat 83a,83b], [Lenat et al. 79]); not being endowed with common sense, an AI system is quite likely to generate terms as absurd as 'grue', and thus we need a theory to guard against inductions using them and a theory to help avoid their generation. At a more basic level, we wish to avoid calling all Italians Giuseppe.

II HIGHER-LEVEL REGULARITIES

The fundamental idea which we aim to expound and formalize is that an inductive generalization can be confirmed or disconfirmed, not only by the observation of its own instances or counter-examples, but also by the observation of other, higher-level regularities in the world. Naturally, these regularities will be based on other instances and, in turn, on other regularities. The general idea is to bring our outside experience to bear on whether to accept a given rule. It is extremely rare for inductions to be performed *in vacuo*. In the case of the traveller in Italy, the generalization that all Italians speak Italian is supported by the more general regularity that, within any given country, most people tend to speak the same language; on the other hand, Giuseppe is not assumed to be the name of all Italians because of the higher-level regularity that almost all social groups use a variety of names. Assuming that emeralds are grue contradicts the general rule that intrinsic properties of objects don't change, particularly not over a whole class and particularly not in response to some absolute time point (as opposed to a time point related to each individual). Some philosophers have objected to the use of such properties as grue in inductions on the grounds that they are intrinsically disjunctive ([Sanford 70]), not ostensively definable ([Salmon 74]), positional and non-qualitative ([Barker & Achinstein 60]) and epistemologically inferior ([Swinburne 73]). But to the little-known species of bond-weevil that lives exclusively on unmatured, fixed-date, treasury bonds, properties such as 'grue' will seem perfectly natural and useful. A theory of induction cannot, therefore, rest on 'intrinsic' properties of the induced rule, but on its relation to the sum of our knowledge of the universe.

In this paper, we will concentrate on confirmatory, rather than disconfirmatory, regularities. Our proposal is that each such regularity corresponds to a universally quantified proposition which, if taken as literally true, would be sufficient to deductively imply the base-level generalization we are attempting, given its observed, positive instances. Furthermore, if the higher-level regularity is to provide additional confirmation, it must have positive instances, preferably a large number, which are *not* instances of the base-level rule. This is the *external evidence* requirement. In a formal system, therefore, the higher-level regularities have the desired *syntactic* relationship to the base-level rule (see the discussion of the syntactic requirement in the Introduction). The higher-level regularities, in turn, may be confirmed by regularities at a still higher level, until ultimately we have to give in to the necessity to do simple enumerative induction. In essence, therefore, we are trying to bring deduction to the aid of induction as far as possible, as a means of allowing our world knowledge to influence our inductive processes.

In the remainder of this paper, we describe the following steps in the process of building a theory of induction:

1) Construction of the space of possible classes of higher-level regularities.
2) Searching the space for interesting classes.
3) Analyzing the results of the search.
4) Applying the results.

III CONSTRUCTING THE SPACE OF HIGHER-LEVEL REGULARITIES

For any particular induction, we can often think of some higher-level rule, derived from our experience, which either confirms or denies it, as in the Italian case. In order to automate this process, we need to elucidate the formal relationship between the base-level and the general rule. We must also endeavour to identify all such classes of general rules, in order that

1) We can take into account all the applicable higher-level rules already known.

2) We can perform further inductions to decide if a potentially relevant higher-level regularity actually holds.

As mentioned above, the higher-level rule, if literally true, should form part of a deductive argument, together with the base-level instances, leading to the affirmation of the base-level rule. Our approach is therefore to construct the space of all possible deductive arguments that lead to the base-level rule as their conclusion. The construction is schematic, i.e., we use generalized predicate schemata P and Q as antecedent and consequent, and the results we are looking for are thus schematic classes of regularities, such that when faced with the task of confirming a given induction, we can instantiate the schematic rule appropriately and use it for steps 1) and 2) given above.

In order to maintain completeness, we construct all resolution proofs of the base-level rule, given the instances. In describing how to do this, we will use rules with unary predicates, for simplicity. As we show below, this results in an overly-restricted space of first-order regularities; this restriction is relieved by using binary predicates. The simplest schematic rule is $\forall x[P(x) \Rightarrow Q(x)]$; we must find those sets of facts which, when combined with the instances and the negation of the rule, lead to a contradiction.

We thus begin with the negation of the rule, which, in clausal (CNF) form, is

$P(a)$

$\neg Q(a)$

for some skolem constant a; we then ask what other facts could be added to lead to a contradiction. To cut a long story short, the only interesting fact we can add is the rule itself, written $\neg P(x) \vee Q(x)$ in CNF. Thus our task becomes that of finding all sets of facts which can be resolved, together with the instances, to leave the base-level rule as the only remaining clause. Since a resolution step removes two complementary literals, our reverse resolution algorithm takes as input the current state of the database, then generates all possible new pairs of complementary literals (and finds all possible existing clauses to which they could be added), such that if the literals were resolved the database would be left in the current state. The literals we introduce can contain any existing predicate, variable or constant, or include a new one of each (designated R_i, y_i, b_i respectively; we choose not to include function symbols in our language). * Thus two possible 'parent' databases of the database containing just the clause $\neg P(x) \vee Q(x)$ are

$\neg P(x) \vee R_1(x)$ and $\neg R_1(y_1) \vee \neg P(x) \vee Q(x)$

$\neg R_1(x) \vee Q(x)$ $\quad R_1(b_1)$

As one might suspect, the space is quite large (in fact, doubly exponential): the base-level database $\neg P(x) \vee Q(x)$ has 20 possible parents; at the second level the average database has around 350 parents. Although we will not discuss them here, our implementation therefore includes a number of pruning heuristics which keep the search manageable without losing any of the interesting points in the space. Another modification is to introduce the instances in a 'macro-move': for the i^{th} instance we add the literals $P(a_i)$ and $Q(a_i)$ as separate clauses, along with their complementary literals attached to other parts of the database, all in one step.

IV SEARCHING THE SPACE

So far we have given a somewhat simplistic picture of the space of regularities we have constructed. As soon as we start searching it, we realize that many of the regularity classes are simply not plausible; that is, they fail to correspond to any possible regularity in the actual world. Unfortunately, this is a hard condition for any machine with limited experience to recognize. For this reason, we currently use a human evaluator for nodes in the space, so that the machine follows paths that the author thinks promising. As a preliminary measure, this has been quite successful; however, to attain our goal of a world-independent theory, and to explore more of the space, we also need to investigate how a machine can recognize that a given class of regularities is uncommon in the world of its experience. It is intended that such a capability be built into the prototype system we have constructed; for our world-knowledge, we will use the broad, common-sense knowledge base of the CYC project ([Lenat et al. 86]). Inasmuch as this knowledge base corresponds to our actual world, this will also constitute empirical research into the actual structure of high-level commonsense knowledge.

It is important for our purposes that the causal structure of the world be such that there are really only a few important classes of regularities. If this were not the case, then whenever we wished to confirm an induction it would be necessary to examine a large amount of potentially relevant information, and to perform a large amount of cumulative detection to maintain the currency of the stock of regularities. Our results so far indicate that there are grounds for optimism, at least in the real world.

V RESULTS

The following subsections describe the general classes of regularity which have been identified after searching the space, with the help of some additional thought. We start with the unary space to illustrate its restrictions, and then move to the binary space. In each section we give the schematic, logical form of the regularity, display the deductive argument leading to the base-level rule, and give an example. Some of these classes were already known to us; others were quite unexpected (sections A and E), although perhaps obvious in retrospect. We are thus convinced of the usefulness of an automatic, (semi-)exhaustive generator of classes of deductive arguments.

A. Rules with a more general left-hand side

The simplest class of higher-level regularities consists of rules with the same consequent as the base-level rule but a weaker (more general) antecedent. Thus the rule

$\forall x[R(x) \Rightarrow Q(x)]$ (where $\forall x[P(x) \Rightarrow R(x)]$)

is sufficient to imply the base-level rule $\forall x[P(x) \Rightarrow Q(x)]$ directly.

Examples:

a) "All social groups use a variety of names" confirms "All nations use a variety of names."
Here $P = Nation$, $Q = NameVariety$, $R = SocialGroup$.

b) "All things made of (a certain type of) beryl are green" confirms

* The exact details of the algorithm used are not important here; one can imagine constructing a simple PROLOG predicate *resolve(Clause1, Clause2, Clause)* which suceeds iff Clause is the result of resolving Clause1 and Clause2; we then invoke the predicate with only the Clause argument instantiated.

"All emeralds are green."
Here $P = Emerald$, $Q = Green$,
$R = MadeOfBeryl$.

Because R is more general than P, the rule $\forall x[R(x) \Rightarrow Q(x)]$ can be confirmed by many instances which are not P; thus, if we have the appropriate data, it becomes easier to prove the more general rule than the more specific (base-level) rule.

This class of confirmation has two apparently distinct interpretations. On the one hand, a) is 'empirical' in flavour: by observing lots of other social groups, we add plausibility to the base-level rule, but no explanation is offered. On the other hand, b) is causal in flavour, offering the beginning of an explanation. Two important points to note here:

- No observations of positive instances of the base-level rule are required.
- The 'explanation' type of support for the generalization is the starting-point for explanation-based generalization [Mitchell et al. 86], which also has no logical need for an instance of the proposed generalization; we can extend the basic principle by adding further intermediary concepts, for example

  $$P(x) \Rightarrow R(x) \quad R(x) \Rightarrow S(x) \quad S(x) \Rightarrow Q(x)$$

  where S is 'reflects light of wavelength 550 nm'. The process of explanation-based generalization uses exactly such a detailed, non-operational theory and compiles it into such useful encapsulations as "all emeralds are green" and "never run outside during an earthquake".

B. Decision rules

The only other simple regularity we have found so far in the unary space takes the form

$$\forall xy[P(x) \wedge P(y) \wedge Q(y) \Rightarrow Q(x)]$$

which can also be written as

$$\forall x[P(x) \Rightarrow Q(x)] \;\vee\; \forall x[P(x) \Rightarrow \neg Q(x)].$$

In [Davies 85] these are called *decision* rules, because P decides the truth of Q. With one instance described by $P(a_1) \wedge Q(a_1)$, the base-level rule becomes deductively justified.
Example:

"Either all cars in Japan drive on the left, or they all drive on the right."

Once we see one car driving on the left, we know that all cars in Japan drive on the left. While it seems true that we can know this decision rule without having been to Japan, in fact it has no confirming instances that are not also instances of the base-level rule. Thus it does not satisfy the external evidence requirement. We actually believe it as a result of a further generalization; if we restrict ourselves to formulae with only unary predicates, we must express this as a second-order regularity, by quantifying over the country predicate P:

$$\begin{aligned}&\forall P[NationalityPredicate(P) \Longrightarrow\\ &\quad \forall xy[P(x) \wedge LeftDriver(x) \wedge P(y)\\ &\qquad \Longrightarrow LeftDriver(y)]]\end{aligned}$$

We will see in the next subsection that this awkward formulation is turned into a first-order sentence by using binary predicate schemata.

C. Direct generalizations using binary predicates

As noted above, using only unary predicates limits the richness of the hierarchy of regularity classes; this limitation is eased when we use binary predicates. The base-level rule that we are now trying to confirm is written $\forall x[P(x,b) \Rightarrow Q(x,c)]$, where b and c are constants. In the unary space, the only interesting database that refutes the negation of the base-level rule was the rule itself. With binary predicates, we also have the following three 'variabilization' generalizations:

$$\forall xy[P(x,y) \Rightarrow Q(x,c)]$$

$$\forall xz[P(x,b) \Rightarrow Q(x,z)]$$

$$\forall xyz[P(x,y) \Rightarrow Q(x,z)].$$

D. More general rules using binary predicates

The binary equivalent of the unary formulae for rules with more general antecedents is

$$\forall x[R_1(x,a_1) \Rightarrow Q(x,c)]$$
where $\quad \forall x[P(x,b) \Rightarrow R_1(x,a_1)]$.

Thus the rule "things made of beryl are green" is expressed as

$$\forall x[Material(x, Beryl) \Rightarrow Colour(x, Green)]$$

The normal type of causal argument introduces a chain of intermediate predicates R_i using appropriate linking constants a_i.

A simple generalization relationship between P and R can also be used:

$$\forall x[R_1(x,b) \Rightarrow Q(x,c)]$$
where $\quad \forall xy[P(x,y) \Rightarrow R_1(x,y)]$.

E. Determination rules

The binary equivalent of a decision rule is called a *determination*, a form which captures a very common and useful type of regularity. The form

$$\forall wxyz[P(x,w) \wedge P(y,w) \wedge Q(y,z) \Rightarrow Q(x,z)]$$

together with one instance described by $P(a,b)$, $Q(a,c)$ is sufficient to guarantee the base-level rule.
Example:

If $Nationality(x,w)$ means "x has nationality w", and $Language(x,z)$ means "x speaks language z", then the determination

$$\begin{aligned}&\forall wxyz[Nationality(x,w) \wedge Nationality(y,w)\\ &\qquad \wedge Language(y,z) \Rightarrow Language(x,z)]\end{aligned}$$

means "Nationality determines Language", since it requires that any two people with the same nationality must speak the same language. With the observation of Giuseppe, an Italian speaking Italian, this gives us the base-level rule "All Italians speak Italian".

Two important points to note:

- Decision and determination rules find a common expression in the extension of predicate calculus described in [Davies and Russell 86], which also shows this form of regularity to be the necessary background knowledge for the successful use of analogical reasoning. We define a new connective, representing the determination relationship as $P(\underline{x}, \underline{w}) \succ Q(\underline{x}, \underline{y})$.
- Determinations also provide a valid form of single-instance generalization which actually utilizes information contained in the instance in forming the generalization. This contrasts with the explanation-based generalization (EBG) technique which simply uses the in-

stance as a focus, assuming that the domain theory is already strong enough to prove the base-level rule. A corollary of this is that, by taking information from the instance, we can build a more powerful single-instance generalization system, in the sense that we can perform the generalization with a weaker domain theory. For example, using the determination "Nationality determines Language", and one instance of an Italian, we predict that all Italians speak Italian; for an EBG system this would require a theory which could predict an entire language (vocabulary, grammar and all) from facts about a nation — needless to say, no such theory is available.

F. Extended determinations

The regularity classes given above are sufficient to guarantee the generalization from no instances or from one instance. Yet quite often we find that one instance is not quite satisfying, but after several confirmations we are happy. One way to account for this is to postulate that the appropriate determination is only weakly supported, so that we need the extra instances to convince ourselves. A different way is to extend the search direction already taken to reach determination, by adding further instances:

$$\forall w, x, y_1, \ldots, y_n, z[P(x,w) \wedge P(y_1,w) \wedge Q(y_1,z) \wedge \\ \ldots \wedge P(y_n,w) \wedge Q(y_n,z) \Rightarrow Q(x,z)]$$

together with n instances described by

$$P(a_1,b),\ Q(a_1,c)\ \ldots\ P(a_n,b),\ Q(a_n,c)$$

is sufficient to guarantee the base-level rule

$$\forall x[P(x,b) \Rightarrow Q(x,c)].$$

The meaning of the extended determination (we might call it determination^n) is clearly seen if we rewrite it:

$$\forall w, y_1, \ldots, y_n, z[P(y_1,w) \wedge Q(y_1,z) \wedge \ldots \\ \wedge P(y_n,w) \wedge Q(y_n,z) \Rightarrow \\ \forall x[P(x,w) \Rightarrow Q(x,z)]]$$

Roughly this can be interpreted as follows "All enumerative inductions from n instances, with P as antecedent and Q as consequent, succeed." This regularity can be confirmed by a history of such successful inductions, and thus the induction in question, $\forall x[P(x,b) \Rightarrow Q(x,c)]$ becomes justified.

As an example, consider again the case of inducing the rule "all emeralds are green", given n green instances. Formally, we write this as

$$\forall x[JewelType(x, Emerald) \Rightarrow Colour(x, Green)].$$

Now many jewel types are not uniform in colour (diamonds, for example, come in black, yellow, blue, pink and white) so the determination "jewel type determines colour" does not hold and we cannot perform a single-instance induction. However, as we explain below, the extended determination *does* still hold, so the n-instance induction is justified.

If we have successfully induced the rules "all sapphires are blue", "all rubies are red", "all amethysts are purple" from collections of instances, then these will be positive instances of the extended determination, so it will be well-confirmed. But in the case of classes such as diamonds, the left-hand side of the extended determination isn't satisfied, since it is unlikely that n instances of a variegated class are all the same colour; thus diamonds are not a disconfirming instance of the extended determination, and it remains well-supported.

If, on the other hand, the Colour predicate admitted arguments like '$grue_{2086}$' (green until 2086, blue thereafter), then the extended determination would have disconfirming instances, since the left-hand side would be satisfied by colours such as $grue_{1972}$ but the universal on the right-hand side would be false.

It is important to note that extended determinations are actually much *weaker* than determinations, and we basically expect them to be satisfied, more or less, for any 'reasonable' P and Q.

VI COMPARISON WITH GOODMAN'S THEORY OF PROJECTIBILITY

Goodman's theory of induction has been the most influential contribution to the field in recent times. We will therefore take the time to briefly outline his theory here, and then re-express it in our terms.

Goodman defines the act of *projection* as the assumption of a general rule from some collection of instances; a rule is *projectible* if this can be done legitimately. The last part of his excellent book, "*Fact, Fiction and Forecast*" ([Goodman 83], first published 1955) is devoted to an attempt to elucidate the criteria for deciding projectibility.

In this theory, rules derive projectibility from three sources:

1) the earned *entrenchment* of the predicates involved;
2) the *inherited* entrenchment which the predicates derive from their *parent predicates*;
3) the projectibility of their *overhypotheses*.

We define these terms below.

A. Entrenchment

Goodman's principal requirement for the projectibility of a rule $\forall x[P(x) \Rightarrow Q(x)]$ is that the predicates P and Q be *well-entrenched.* A predicate P becomes well-entrenched as an antecedent as a result of frequent past projections of other rules with P as antecedent; similarly for Q as consequent. Thus 'green' is well-entrenched, whilst 'grue' is not.

B. Parent predicates

The notion of a parent predicate is used in defining both inherited entrenchment and overhypotheses. A predicate R is a parent of S iff

1) R is a predicate applying to classes of individuals.
2) Among the classes to which R applies is the extension of S. Thus 'uniform in colour', which applies to any group of individuals all of the same colour, is a parent of 'green'. Similarly, 'type of jewel' is a parent of 'emerald'.

C. Inherited entrenchment

A predicate inherits entrenchment from its parent predicates. Thus if 'uniform in colour' is well-entrenched, 'green' derives further entrenchment from it.

D. Overhypotheses

An overhypothesis of $P \Rightarrow Q$ is a rule $R \Rightarrow S$ such that R is a parent of P and S is a parent of Q. Thus an overhypothesis of "all emeralds are green" is "all types of jewels are uniform in colour". If the overhypothesis is projectible, this adds to the projectibility of its underhypothesis. Here, for example, both R and S are reasonably entrenched, and the overhypothesis is fairly well supported, e.g. by "all sapphires are blue", "all rubies are red". A given rule can have many overhypotheses, and each may in turn be supported in turn by further overhypotheses at the next level.

E. Analysis

We will now attempt to analyze Goodman's theory in our terms. By formalizing each of his notions, we can fit them into the general framework of the confirmation of rules by higher-level regularities.

The entrenchment of a predicate P corresponds approximately to an observed second-order regularity of the form

$$\forall Q \forall x_1 \ldots x_n[[P(x_1) \wedge Q(x_1) \wedge \ldots \wedge P(x_n) \wedge Q(x_n)] \\ \Rightarrow \forall x[P(x) \Rightarrow Q(x)]]$$

which bears close resemblance to the definition of extended determination given above. The difference is that because Goodman is working exclusively with unary predicates, he is forced to quantify over the predicate Q (in defining the entrenchment of P) in order to satisfy the external evidence requirement, thus requiring that P be a successfully projected predicate *regardless* of the consequent Q. The use of binary predicates allows us to quantify just over their second argument, giving the more fine-grained notion of successful projection of *similar* rules, rather than just rules with the same antecedent.

The notion of a parent predicate is a little tricky to formalize using unary predicates; it would look something like this:

$$A \text{ is a parent of } B \quad \text{iff} \quad \exists S[A(S) \wedge \forall x[x \in S \Leftrightarrow B(x)]]$$

A more natural way to write it is to use a binary predicate:

$$A \text{ is a parent of } B \quad \text{iff} \quad \forall x[B(x) \Leftrightarrow A(x, B)]$$

which amounts to *reifying* B. For example, we write

$$\forall x[Emerald(x) \Leftrightarrow JewelType(x, Emerald)]$$

Viewed in this light, an overhypothesis is essentially a determination.

Clearly, there is a great deal of overlap in the two approaches. There are, however, some slight differences in emphasis, stemming mainly, one may conjecture, from the differing requirements of philosophy and artificial intelligence.

- Goodman is trying to systematize human practice; he does not attempt, for example, to *justify* the entrenchment criterion. When written formally, we see entrenchment (and the other notions) as codifications of higher-level regularities, which push back the inevitable point at which we must simply appeal to an unjustifiable, naked principle of enumerative induction. (As is pointed out in [Quine & Ullian 70], in the human case we may be able to push it back far enough such that the evolutionary process itself may be 'credited' with performing such inductions.) The main commonality of the two theories, and the revolutionary aspect of Goodman's work, is that we no longer have to make such an appeal within the base-level induction itself.
- In Goodman's theory, predicates derive entrenchment from actual past projections, taking the form of (not necessarily spoken) linguistic utterances and corresponding to projections performed in the history of the culture rather than just the individual. This is essentially a psychological theory about exactly what evidence humans take into account in making new projections. In our approach, we try to identify all the evidence that should logically be taken into account, which may entail making further inductions 'on demand' as well as noticing past inductions.
- Because we use binary predicates and an exhaustive generator, we are able to produce a much richer hierarchy of 'overhypotheses'. Both theories, however, rely on the existence of a rich taxonomic vocabulary to facilitate expression of the desired regularities. This leads us naturally into a study of the relation between language and induction.

VII REPRESENTATION AND INDUCTION

An implicit hypothesis of Goodman's theory is that everyday terms will tend to be well-entrenched, since otherwise they would drop out of use. (He states (p. 97) that "entrenchment and familiarity are not the same . . . a very familiar predicate may be rather poorly entrenched," but gives no examples.) The key idea behind analyzing this hypothesis is to understand the process by which terms become familiar parts of the language. If we can capture the conditions under which new words are acquired, then we can give a semantics to the *presence* of a word in our language, as well as to the word itself. * Thus the fact that green is a primitive attribute in our language, as well as being a physiological primitive of our observation apparatus, suggests that greenness is a commonly-occurring property in the world, and, more importantly, that greenness is a *good predictor* for various other properties, such as whether something is animal or vegetable, ripe or unripe. If we limit our acquisition and retention of terms to those which manifest such useful properties, then we are guaranteed that familiar terms will tend to be entrenched, and thus that rules using them wili be projectible. The language-evolution aspect of this idea finds strong echoes in the theory of induction given in [Christensen 64]; the reflection of properties and regularities of the world in our neurological development is one of the principle themes of Roger Shepard's work, described in [Shepard 84, 86]. Although we have barely scratched the surface of the enormous topic of the interrelationship of language, representation and learning, it seems that the analysis of the semantics of the presence of words in a language, via the analysis of the processes of acquisition and retention, may be a profitable approach.

VIII APPLICATIONS

We will first describe how we propose to build systems utilizing the ideas given above; we will then discuss possible applications to some induction projects, past and present.

The scenario we envisage is that of an autonomous intelligent agent engaged in the continuous process of investigating its environment and attempting to fulfil its various goals. The system may need to assess the degree of confirmation of a proposed rule for one of three reasons:

1) it needs a rule for concluding some goal, and has none available;
2) it has some theoretical reasons for believing the rule plausible;
3) it has noticed that the rule is empirically plausible.

* Rendell, in [Rendell 86], talks about the "semantics of the constraint imposed by the language" as part of an attempt to understand the bias inherent in version-space systems (the ungrounded premise to which we alluded earlier); this is another aspect of the same idea.

To evaluate the proposed rule, the system performs the following tasks:

- Assess the direct empirical support for the rule; if necessary, this may involve experimentation.
- Instantiate the known classes of higher-level regularity so that they apply to the rule in question; if the system already knows the degree of confirmation of the instantiated regularities, take that into account; if not, call the evaluation procedure recursively to compute their confirmation.
- Repeat the same process for any plausible competing hypotheses.

If the proposed rule is well-supported by its higher-level regularities, and clearly better than any conflicting hypothesis, then it can be adopted (subject to revision).

From our investigations to date in the space of regularities, it seems that we can capture most of the relevant information using just three basic classes: simple implicative rules, determinations and extended determinations. These seem to provide the justification for the basic types of argument on common use. As mentioned above, as long as there are a small number of types it is reasonable to build specialized 'regularity-noticing' demons to spread the computation load, rather than using 'lazy evaluation'. The higher-level rules we thus accumulate are also useful for suggesting new, plausible base-level rules.

Our proposed architecture seems closest to that of AM and EURISKO ([Lenat 76], [Lenat 83a,83b]), which actively performs experiments in order to confirm its conjectures inductively. EURISKO can be said to use higher-level regularities of a sort, since it has a heuristic which essentially leads it to consider conjectures similar to those which have already proven successful. Recalling the basic task of inferring facts from a mass of ground data, it is clear that when we add the ability to recognize a new class of higher-level regularities we actually expand the set of inferences the system can make. Most inductive systems in AI use only simple, associative regularities. We therefore hypothesize that with the degree of synergy afforded by the addition of multiple layers of regularities, EURISKO's performance can be considerably enhanced.

A system which uses theoretical (causal, explanatory) support as well as direct empirical support for its proposed rules is described in [Doyle 85]. In the light of the theory given above, we would argue that there are forms of further, indirect empirical support which are in no sense causal, yet offer more power than the simple 'associationist' approach. Other systems which conduct large-scale inductive investigations are the RX system ([Blum 82]), and UNIMEM/RESEARCHER ([Lebowitz 86]); the same arguments apply in these cases.

IX CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

We have shown that the requirement for a theory of induction is not that it render enumerative induction valid, but that it elucidate the way in which the plausibility of an induction is affected by the presence of further evidence, distinct from its direct positive and negative instances. The relationship between the direct and indirect evidence is a formal one, as required, and we have given a method for identifying all general classes of such evidence. We have constructed a system which applies the method to discover some novel and, we believe, important classes of regularity. The result of the synergistic interplay of induction and deduction is that we can now distinguish plausible from spurious inductions, and can maximize the usefulness of the observational knowledge a system possesses. The 'punchline' is simply this: *the more classes of regularity a system is equipped to observe, the more inferences it can make from a given collection of data.*

A major weakness which we would like to address is that the theory as described only allows first-order regularities. Although we glossed over the point in the exposition above, an extended determination need not use only an exact number n for all its inductions — n really just means 'many', and this is how it will be implemented in the real system. The model of analogy by similarity in [Russell 86] suggests that there may be other useful non-first-order regularities, for example in the definitions of natural kinds ([Rosch 78]) and in the distributional variation of attribute values in a population ([Medin & Smith 84]). At present it is not clear how to cope with these problems.

Potentially fruitful areas for further investigation include:

- studying the interaction of language and induction via the semantic analysis of the process of representational evolution;
- empirical experiments to establish what are the useful, commonly-occurring classes of regularity in any given world;
- quantification of the contributions of higher-level regularities to a base-level rule, especially regularities with less than 100% confirmation;
- construction of robust systems, using the principles outlined above, that are able to acquire, organize and use effectively knowledge of a complex environment, even in the absence of any *a priori* knowledge of the environment; although such systems seem somewhat beyond our present abilities, it is hoped that we have begun to dismantle one of the theoretical barriers to their creation.

ACKNOWLEDGEMENTS

I would like to thank my advisors Doug Lenat and Mike Genesereth and colleagues Todd Davies (SRI), Devika Subramanian and David Wilkins (Stanford) and all the members of the Logic Group of the Stanford Knowledge Systems Laboratory for fruitful discussions, constructive criticism and moral support.

References

[Barker & Achinstein 60]
Barker, S. F. & Peter Achinstein. "On the New Riddle of Induction". In *Philosophical Review*, vol. 69, pp. 511-22; 1960.

[Blum 82]
Blum, R. L. *Discovery and representation of causal relationships from a large time-oriented clinical database: the RX project.* Ph. D. thesis, Stanford University, 1982.

[Christensen 64]
Christensen, Ronald. *Foundations of Inductive Reasoning.* Berkeley: 1964.

[Davies 85]
Davies, Todd. *Analogy.* Informal Note No. IN-CSLI-85-4,

Center for the Study of Language and Information, Stanford University; 1985.

[Davies & Russell 86]
Davies, Todd & Stuart Russell. *A Logical Approach to Reasoning by Analogy.* Stanford CS Report (forthcoming) and Technical Note 385, AI Center, SRI International; June, 1986.

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

[Doyle 85]
Doyle, Richard J. *The Construction and Refinement of Justified Causal Models through Variable-level Explanation and Perception, and Experimenting.* Ph.D. thesis proposal. Massachusetts Institute of Technology; 1985.

[Goodman 46]
Goodman, Nelson. "A Query on Confirmation". In *Journal of Philosophy*, Vol. 43, pp. 383-5; 1946.

[Goodman 83]
Goodman, Nelson. *Fact, Fiction and Forecast*, 4th edition. Cambridge, MA and London: Harvard University Press; 1983. (First published 1955).

[Holland et al. 86]
Holland J., Holyoak K., Nisbett R. & Thagard P. *Induction: Processes of Inference, Learning and Discovery.* In press.

[Hoppe]
Hoppe, Arthur. "Our perfect economy". In *San Francisco Chronicle.* San Francisco: Date unknown.

[Lebowitz 86]
Lebowitz, Michael. "Concept Learning in a Rich Input Domain: Generalization-based Memory". In Ryszard S. Michalski, Jaime G. Carbonell & Tom M. Mitchell (Eds.), *Machine Learning: an Artificial Intelligence Approach; Volume II.* Los Altos, CA: Morgann Kaufmann, 1986.

[Lenat 76]
Lenat, D. B. *AM: An artificial intelligence approach to discovery in mathematics as heuristic search.* Ph.D. thesis, Stanford University, 1976.

[Lenat 83a]
Lenat D. B. "Theory formation by heuristic search. The nature of heuristics II: Background and Examples". In *Artificial Intelligence*, Vol. 21, Nos. 1,2; 1983.

[Lenat 83b]
Lenat D. B. "EURISKO: A Program That Learns New Heuristics and Domain Concepts. The Nature of Heuristics III: Program Design and Results". In *Artificial Intelligence*, Vol. 21, Nos. 1,2; 1983.

[Lenat et al. 79]
Lenat, D. B., Hayes-Roth, F. and Klahr, P. *Cognitive Economy.* RAND Technical Report No. N-1185-NSF. Santa Monica, CA: The RAND Corporation; 1979.

[Lenat et al. 86]
Lenat D., Mayank P. and Shepherd M.. "CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks." *AI Magazine* Vol. 6 No. 4; Winter 1986.

[Medin & Smith 84]
Medin D. L. & Smith E. E. "Concepts and Concept Formation." In *Annual Review of Psychology* Vol. 35; 1984.

[Michalski 83]
Michalski R. S. "A Theory and Methodology of Inductive Learning." In *Artificial Intelligence*, Vol. 20, No. 2; Feb 1983.

[Mitchell 78]
Mitchell, Tom M. *Version Spaces: an Approach to Concept Learning.* Ph.D. thesis, Stanford University, 1978.

[Mitchell et al. 86]
Mitchell, T. M., Keller R. M., Kedar-Cabelli S. T. "Explanation-based Generalization: a Unifying View". In *Machine Learning Journal* Vol.1 No. 1; 1986.

[Quine & Ullian 70]
Quine W. V. & Ullian J. S. *The Web of Belief.* New York: Random House; 1970.

[Rendell 86]
Rendell, Larry. "A General Framework for Induction and a Study of Selective Induction." In *Machine Learning Journal* Vol. 1 No. 2; 1986.

[Rosch 78]
Rosch, E. "Principles of categorization". In *Cognition and Categorization*, Rosch E. and Lloyd B. B. (Eds.). Hillsdale: Lawrence Erlbaum Associates; 1978.

[Russell 86]
Russell, Stuart J. "A Quantitative Analysis of Analogy by Similarity". In *Proceedings of the National Conference on Artificial Intelligence.* Philadelphia: AAAI; 1986.

[Salmon 74]
Salmon, Wesley. "Russell on Scientific Inference". In G. Nakhnikian (Ed.), *Bertrand Russell's Philosophy.* New York: Barnes and Noble; 1974.

[Sanford 70]
Sanford, David H. "Disjunctive Predicates". In *American Philosophical Quarterly*, Vol. 7, pp. 162-70; 1970.

[Shepard 84]
Shepard, Roger. "Ecological Constraints on Internal Representation: Resonant Kinematics of Perceiving, Imagining, Thinking and Dreaming". In *Psychological Review* Vol. 91, No. 4. October, 1984.

[Shepard 86]
Shepard, Roger. *Mind and World.* Forthcoming.

[Swinburne 73]
Swinburne, Richard. *An Introduction to Confirmation Theory.* London: Methuen; 1973.

[Utgoff 84]
Utgoff P. E. *Adjusting Bias in Concept Learning.* Ph.D. thesis, Rutgers University, 1984.