Volume 5
Number 3
1992

International Journal of Expert Systems

Research and Applications

SPECIAL ISSUE:

Inference and Machine Reasoning

GUEST EDITOR:
Alberto J. Cañas
University of West Florida

JAI PRESS
Greenwich, CT • London, England

A Model for Abductive Problem Solving Based on Explanation Templates and Lazy Evaluation

CARL STERN
Department of Computer Science
University of New Mexico

GEORGE F. LUGER
Department of Computer Science
The University of New Mexico

ABSTRACT: A new model for abductive problem solving is proposed. The set cover and logic-based approaches to abduction are first criticized. The set cover approach cannot handle situations involving causal interactions or causally structured event sequences. The logic-based approach is semantically weak and computationally expensive. The *deductive-nomological* theory of explanation, a theory often used to support logic-based abduction, is examined. We show that the requirement of causal or deductive completeness in explanations is naive and unrealistically restrictive.

The account of abductive problem solving offered to replace the logic-based model is based on explanation templates and a form of lazy evaluation. Explanation schemas or templates organize the initial problem description and guide the interpretation of the evidence. Each template has its own principles of interpretation associated with it, its own set of data-signs which it focuses upon, and its own procedures for evaluating those signs. Explanation templates are hierarchically organized from abstract to concrete, allowing the problem solver to ascribe a type of causal mechanism to a situation before the details of the explanation are investigated. We describe a diagnostic example, i.e. discrete semiconductor failure analysis, which illustrates the use of explanation templates in abductive problem solving. Finally we discuss a computational model which incorporates many of the features described.

1. INTRODUCTION

In recent years AI researchers have recognized the limitations of classical deductive logic. It has been observed that sound deductive inference accounts for only a relatively small part of human problem solving. Human problem solvers frequently function with incomplete information and uncertain or imprecise domain theories, yet manage with remarkable regularity to find effective solutions. The desire to model the inexact components of human

International Journal of Expert Systems Pp. 249–265, Volume 5, Number 3, ISSN 0894-9077

problem solving has led to work in non-monotonic reasoning, fuzzy logics, Bayesian belief networks, etc.

Research on abduction also arises from the effort to understand the inexact or unsound[1] side of effective human problem solving. Abduction can be defined as *reasoning to the best explanation* (Thagard, 1989). This is generally construed as the use of some form of principled reasoning to discover the causes for a given observation or event (Levesque, 1989).

There are several reasons for pursuing research on abductive reasoning. Abductive reasoning plays an important role in medical diagnosis, equipment maintenance and repair, chemical analysis, and other forms of investigative problem solving typically modeled by expert systems. The explicit study and formalization of abductive reasoning may thus provide reasoning tools for a significant class of expert systems.

The construction of explanatory hypotheses is a pervasive feature of common sense reasoning and natural language understanding as well. Our understanding of the world at the common sense level is, to a great extent, knit together by reasoning from effects to back to causes. The ability to 'infer' an unobserved action or characteristic of an object from its observed effects supports the formation of expectations about the future based on the regularities of the past (Peirce, 1955). In natural language understanding, researchers have argued that abduction plays an important role in extracting implied events and actions from narrative descriptions (Ng & Mooney, 1990).

In making the notion of abduction more precise, we distinguish three things to which this term can refer:

1) a particular act of adopting a revocable belief in some explanatory hypothesis on the grounds that it best explains an otherwise unexplained set of facts;
2) The 'inference rule' warranting such cognitive acts and defining the legitimate form(s) of abductive 'inference';
3) the cognitive mechanism(s) involved in the construction or discovery of explanations.

Most recent accounts of abduction in AI have taken either a set cover or a logic-based approach. Set cover approaches address item 3) alone, whereas the logic-based approaches often address both 2) and 3). In the set cover approach, an abductive explanation is defined as a *covering* of observations by hypotheses, where these covers rest on a binary 'causal' relation $R \subset Hypotheses \times Observations$. (Reggia, 1983) Thus an abductive explanation of a set of facts *S2* is another set of facts *S1* sufficient to cause *S2*. An optimal explanation in this model is a minimal set cover. The weakness of this approach is that it reduces explanation to a simple list of causes. In situations where there are interrelated or interacting causes or where an understanding of the sequence and structure of causal interactions is required, the set cover model proves inadequate.

Logic-based approaches, on the other hand, rest on a more sophisticated notion of explanation. An abductive explanation of some representation E is defined as a set of hypotheses (H) consistent with an agent's background knowledge (B) such that $H \cup B$ logically entails E. On this model, the proof tree for E supplies the causal connections between the explanatory hypotheses H and the explanatory goal E. (In this sense only the proof tree constitutes a *complete* explanation; however, the hypothesis set alone is often loosely referred to as an *abductive* explanation.)

The logic-based definition of abductive explanation suggests a corresponding mechanism for explanation discovery. If the explanatory hypotheses must entail E, then the way to find such explanatory hypotheses is to reason backwards from E. In a Horn clause set, the natural interpretation of this is to start from the conjunctive components of E and backchain from consequents to antecedents. Thus the definition of abductive explanation and the mechanisms for discovering explanatory hypotheses have a similar form:

Abductive Explanation:

Explain $(\beta,\alpha \vdash \beta, B)=\alpha$	if $\alpha \cup B$ is consistent
Backchaining Primitives:	
Abduce $(\beta,\beta\leftarrow\alpha,B)=\alpha$	if $\alpha \cup B$ is consistent
Abduce $(B,\{\beta\leftarrow\alpha,\alpha_i\leftarrow\delta\},B)=\{\alpha_1,..,\delta,..\alpha_n\}$	if $a=\{\alpha_1\wedge\alpha_i..\wedge\alpha_i..\wedge\alpha_n\}$ and $\{\alpha_1,..,d,..a_n\}\cup B$ is consistent

This model appears natural because the conditionals which support backchaining can readily be thought of as *causal laws*, thus capturing the pivotal role which causal knowledge plays in the construction of explanations. The model is also convenient because it dovetails nicely with something which the AI community already has a great deal of experience at: the computational modeling of deduction and backchaining.

As simple, precise, and convenient as the logic-based account of abduction is, it has nonetheless some important shortcomings. Logic-based computational models of abduction encounter two sorts of difficulties. The first involves computational complexity. Logic-based models are computationally expensive, requiring in the worst case exponential work. The second has to do with semantic weakness. The entailment relation provides an inadequate basis for drawing important semantic distinctions, that is, for distinguishing good (edifying, coherent, plausible) explanations from poor (trivial, tangled, implausible) explanations.[2] Thus the logic-based approach is capable of generating a very large set of candidate explanations from which a few serious ones must be sifted. These two problems stem from a common source. We argue that the excessive computational expense of logic-based models is at least partly a result of their lack of semantic structure.

In our experience with diagnostic problem solving, the logic-based approach is often impractical. In many practical situations the human expert is not capable of meeting such a rigorous standard of explanation (i.e., a deduction of the *explanandum* from the explanatory hypotheses). In those situations where a deductive explanation is required, it is generally not constructed in a single pass but rather incrementally and interactively, with further evidence gathering and validation steps along the way. More often, however, a complete deductive explanation is not required. What is needed is only the identification of a set of responsible elements in a situation, elements whose undesirable effects can be transformed by practical intervention, while other contributory aspects of the situation remain constant. The successful termination of the abductive process is thus defined relative to some practical goal. In all these cases the backchaining model does not accurately reflect the incremental heuristic character of the process by which explanations are actually constructed.

We offer in this paper a model of heuristically informed abduction which preserves some of the strengths of the logic-based approach while fitting better the practical requirements

of many problem solving situations. This model incorporates four features which are widely found in the problem solving approaches of human experts:

1) The use of the explanation schemas – with powerful abstraction mechanisms – to organize the construction of explanations;
2) descriptions (interpretations) which change with shifts in explanatory perspective;
3) evidence gathering informed by currently held explanatory hypotheses;
4) lazy evaluation: a mechanism for filling in the details of an explanation on an as-needed basis, postponing the precise specification of certain elements of a causal mechanism until (and only if) a more precise account of those elements is required.

In section 2 we criticize logic-based approaches to abduction. We examine a recent paper (Selman, 1990) which shows that logic-based abductive reasoning involves an NP-hard "computational core". We examine the *explanation selection* task in logic-based models. We look at mechanisms for integrating considerations of coherence, plausibility, and interest into logic-based abduction. We argue that the separation of the *explanation construction* and the *explanation selection* tasks is a mistake, the result of which is to eliminate important semantic and heuristic supports from the explanation construction process.

In section 3 we explore the *deductive-nomological* theory of explanation. This theory provides important philosophical support for the logic-based account of abduction. We examine the relationship between causality and entailment, challenging the claim that explanations in which the deducibility requirement is not met are necessarily inadequate or incomplete. We argue that acceptable explanations often include implicit place-holders for unspecified causal mechanisms, effectively bypassing the deducibility requirement.

In section 4 we revisit Peirce's treatment of abduction. Abduction, on Peirce's account, involves the discovery of the general pattern or rule to which the instance conforms. It thus encompasses type or pattern-based explanation as well as causal explanation. We use an examination of Peirce's account to illuminate the subtle relationship between the problem description and the eventual explanation. This relationship can be encapsulated in the paradoxical observation that only after a situation is actually explained can one be certain of exactly what facts needed to be explained.

In section 5 we amplify Peirce's view with some general observations about investigative processes. We then describe an example of abductive problem solving involving discrete semiconductor failure analysis. In the context of this example we further elaborate certain essential features of abductive problem solving.

In section 6 we propose a computation model of abductive problem solving with some of the features described in section 5. This model is based on explanation templates and a form of lazy evaluation. We describe how an implementation of explanation templates is used in a semiconductor failure analysis expert system to capture the adaptive, heuristic character of abductive problem solving.

2. THE COMPLEXITY AND SEMANTICS OF "ABDUCTIVE LOGIC"

Selman and Levesque examine the complexity of abduction tasks similar to that computed by an ATMS.[3] The standard proof that the ATMS problem is NP-hard depends on the

existence of queries with an exponential number of explanations. They attempt to bypass the issue of the number of potential explanations by asking whether finding a single or a small number of explanations is also NP-hard. Finding a single explanation given a knowledge base K containing arbitrary clauses is easily shown to be NP-hard by a reduction to SAT (since explanations only exist when K is consistent). But given a Horn clause knowledge base Σ they produce an algorithm that finds a single explanation in O(kn) where k is the number of propositional variables and n is the number of occurrences of literals in Σ. However when restrictions are placed on the kind of explanation sought the problem again becomes NP-hard, even for Horn clause KBs. They show that goal directed abduction (generating an explanation for q that contains a specified proposition p) is NP-hard. They also show that finding an assumption-based explanation (an explanation of q consisting soley of members of some specified assumption set A) is NP-hard. They trace the difficulty of these tasks as well as that of default reasoning in Reiter's Default Logic to a common computational core which they call the Support Selection Task.

One of the more interesting results of their analysis is the fact that adding goals or restrictions to the abduction task actually makes it significantly harder. From a naive viewpoint of a human problem solver this is surprising: a human problem solver might assume that the addition of focus to his/her search would make the task easier. The reason that it makes the abduction task harder in the logic-based model is that it only contributes additional *constraints* to the problem, not additional structure to the problem solving.

Explanation discovery is characterized in the logic-based model as the task of finding a set of hypotheses with certain logical properties. These properties (consistency with background knowledge, entailment of the *explandum*) are meant to capture the *necessary* conditions of explanation, that is, the minimal conditions which a set of explanatory hypotheses must satisfy in order to count as an abductive explanation. Proponents of this approach believe that by adding additional constraints it can be extended to provide a characterization of good or reasonable explanations.

One simple strategy for refining this account is to define a set of unit clauses which are 'abducible', that is, from which candidate hypotheses must be chosen. This allows search to be restricted in advance to those factors that can potentially play a causal role in the chosen domain. Another strategy is to add selection criteria for evaluating and choosing between abductive explanations. Various selection criteria have been proposed (Levesque, 1989). Set minimality prefers hypothesis set *S1* over *S2* if $S1 \subset S2$ (where both are consistent and entail the *explanandum*). The simplicity criterion gives preference to parsimonious hypothesis sets, ones containing fewer unverified assumptions. Both minimality and simplicity can be seen as applications of Ockham's razor. Unfortunately set minimality is of limited power as a pruning tool; it only eliminates explanations which are supersets of existing explanations.

Simplicity alone, on the other hand, is of questionable validity as a selection criterion. It is not difficult to construct examples in which an explanation requiring a larger hypothesis set is preferable to some simpler but shallower one. Indeed, complex causal mechanisms will generally require larger hypothesis sets; however the abduction of such causal mechanisms may well be justified, particularly when the presence of certain key elements of that mechanism have already been verified by observation.

Two other mechanisms for explanation selection are interesting because they take into account not merely properties of the hypothesis set but also properties of the proof tree. Cost-based abduction places a cost on potential hypotheses and also a cost on rules. The total cost

of the explanation is computed on the basis of the total cost of the hypotheses and the cost of the rules used to abduce the hypotheses. Competing hypothesis sets are then compared according to cost. The most natural semantics that can be attached to this scheme is a probabilistic one (Charniak, 1990). Higher costs for hypotheses represent less likely events; higher costs for rules represent less probable causal mechanisms.

Coherence based selection criteria are particularly appealing when the *explanandum* is not a simple (elementary) proposition but a set of propositions. Ng and Mooney (1990) have recently argued that a coherence metric is superior to a simplicity metric for choosing explanations of descriptions occurring in natural language text. They define coherence as a property of the proof graph where explanations with more connections between any pair of observations and fewer disjoint partitions are more coherent. The coherence criterion is based on the heuristic assumption that what we are asked to explain is a single event (action) with multiple aspects. They justify their use of a coherence criterion for natural language understanding in terms of Gricean felicity conditions, that is, the speaker's obligation to be coherent and pertinent. It is not difficult to extend their argument to a variety of other situations. In descriptions of problem solving situations the observations which comprise the initial *explanandum* are brought together because they are believed to be related to the same underlying failure mechanism.

The proposed mechanisms for explanation selection can all be regarded as useful to one degree or another. However they all fail to address certain semantic considerations which humans consider in constructing and evaluating explanations. For this reason, Reiter (1987) speculates that the logic-based approach may be useful in characterizing the *explanation construction* task but argues that extra logical considerations are required for the *explanation selection* task.

To bring into view these elusive semantic considerations we look at the following series of examples:

1) α trivially entails α. Thus α explains α for any proposition α.
2) A circuit has an output which is 0, 1, or undefined. This can be represented $\mathbf{p \vee q \vee r}$. Suppose we are required to explain why the output is stuck on 0. One explanation will be $\{\neg \mathbf{q}, \neg \mathbf{r}\}$.
3) Heating causes copper to expand. This can be represented $\mathbf{p} \supset \mathbf{q}$. The contrapositive is $\neg \mathbf{q} \supset \neg \mathbf{p}$. This can play an unanticipated role in explanation. Suppose we have placed a piece of copper in a vessel over a bunsen burner and it does not heat. One explanation which may be generated is that it does not heat because it does not expand.
4) Two parents who desire a daughter take steps (in accordance with the latest in medical technology) which they believe will cause their child to be female rather than male. Suppose however that their child turns out to be male. Suppose a representation of their background knowledge includes the following semantic information about sons: a son is a male child $(\mathbf{p} \supset (\mathbf{q} \wedge \mathbf{r}))$. Then one explanation for the maleness of their child will be that they had a son.

The problem in Example 1 is that the explanation is circular. We might consider an *ad hoc* rule which eliminates all explanations of α which include α itself. However circular explanations are not always so easily recognized; α may be disguised, for example, by replacing it with some semantically equivalent expression.

Example 2 rests on the *disjunctive syllogism.* It is used to discover a fact by eliminating its alternatives. It involves evidential rather than causal reasoning and is thus inappropriately used in the construction of an explanation.

Example 3 starts with a causal relationship but transforms it into an evidential one by rewriting the implication in the contrapositive form. Both 2 and 3 can be eliminated as problems by limiting representation to Horn clause form.

Example 4 involves use of *analytic* or *tautological* relationships to construct an explanation. This can lead to trivial explanations like the one given. However *analytic* rules can sometimes play an important role in explanation: when expanding an expression α by replacing it with an equivalent expression β allows access (matching) to other causal laws which were expressed in terms of β.

All these examples point out the important role played by certain kinds of general rules or laws in the construction of explanations. As we have seen, when none of the rules upon which an explanation is based are semantically *interesting* then there is a danger that the explanation will be trivial. The difficulty faced by the logic-based approach is that there is no way of guaranteeing semantic significance through the syntactic form of a rule, $\alpha \supset \beta$. The pivotal role played by general rules or laws in the construction of explanations is the focal point of the *deductive nomological* theory of explanation. This theory, described in section 3, offers a much tighter account of the logical form of explanation.

3. THE DEDUCTIVE-NOMOLOGICAL THEORY OF EXPLANATION

Logic-based accounts of abduction frequently claim as their justification a fairly robust philosophical theory, the *deductive nomological* or *covering law* theory of explanation. On this account, an explanation is something with the logical form shown in Figure 1. Here $C_1, C_2, .., C_m$ represent some set of relevant initial conditions, $L_1, L_2, .., L_n$ represent a set of general laws, and E represents the thing to be explained. For the explanation to be valid it is required that the *explanandum* be deducible from the *explanans*, that is, the relevant initial conditions and general laws (Hempel, 1965).

Observe that the presence of the, L_i in this schema is meant to capture the notion that explanation depends on general rules or laws. These covering laws supply the intelligible connection between the conditions C_i and the explanandum E. Thus the type of the explanation can generally be determined from the type of covering laws. In mathematical explanations, the covering laws are axioms or theorems. In morphological explanations, the covering laws are type or species invariants (e.g., copper is conductive, bees gather plant

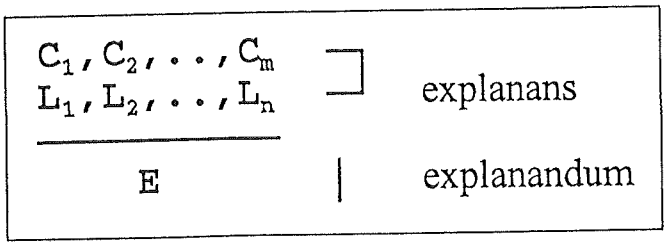

Figure 1. Form of a deductive nomological explanation

pollen). In causal explanations, the covering laws are causal laws. In trivial explanations, the covering laws are vacuous tautologies (e.g., α implies α, a son is a male).

To the degree that the covering laws in an explanation are expressed as truth functional conditionals, they are required to assert deterministic relationships. The rule $\alpha \supset \beta$ is falsified if in some situations α is present and β is not. Frequently, however, explanations rest on weaker relationships. For example, a child's exposure to someone with chicken pox is often used to explain contraction of the disease in spite of the fact that not all exposed children contract chicken pox. Similarly, a weak battery may be used to explain why a car won't start when the car started yesterday with the battery in a nearly identical condition.

These explanations are considered *incomplete* in the *deductive-nomological* account.[4] This way of categorizing explanations is fine as long as incompleteness is not confused with inadequacy. Deductive completeness is a standard that can rarely be satisfied. Generally deductive completeness occurs only in explanations which deal with simple, strictly controlled situations (e.g., in a laboratory or a circuit board). Eliminating all *incomplete* explanations would eliminate most of human problem solving.

Most rules expressing causal relations have vaguely specified application conditions: Under sufficient stress metal fatigue can lead to metal disintegration. In low temperatures a weak battery can cause a car not to start. For many familiar and useful causal mechanisms a problem solver is incapable of specifying precisely and exhaustively the necessary and sufficient conditions under which the causal mechanism is effective.

Incompleteness is not necessarily a practical concern. In many problem solving situations we are interested in identifying the responsible elements of the situation. Responsibility is validated by transforming specific elements to determine whether such a change achieves a desired goal. Even though other elements of the situation may play a contributory role, their incomplete description is not a problem because they remain constant throughout the intervention.

The vagueness and generality of causal rules, in fact serves a useful purpose: organizing and simplifying search. General rules such as under the right circumstances metal fatigue causes structural failure in an airplane' allow the discovery of metal fatigue as a potential cause without sorting through the finer situational details. If and when the hypothesis of metal fatigue becomes plausible because of certain evidence, the precise extent of metal fatigue and the nature of in flight stresses can then be sorted through in a more focused fashion.

One of the weaknesses of the covering law and logic-based accounts is that they do not take into account the hierarchal structure of explanation schemata. Causal or explanatory rules are generally formulated at different levels of abstraction corresponding to the differing focuses and levels of detail required at different points in the explanation process.

For example in medical diagnosis the following hierarchy of explanation schemata may be observed:

1) pathogens cause disease;
2) meningitis (a pathogen induced disease) causes a symptom set S, for example, headache, neuralgia, fever, elevated white count,...;
3) pseudomonas-aeraginosa infection causes symptom set S* (a form of meningitis) in patients with a weakened immune system.

We will present an account of explanation based on hierarchical abstraction in section 5.

that the eyeglasses the priest was wearing was not a significant indicator. One is faced with the paradoxical conclusion that only after a situation is explained can one be certain of precisely what facts needed to be explained.

The logic-based approach does not attempt to model the process of selecting the set of facts needing to be explained. Logic-based theorists seem to overlook the fact that the construction of the problem description is itself an abductive process resting on background or domain knowledge. In the logic-based approach the explanatory goal or *explanandum* is given at the outset and fixed throughout the problem solving process. This does reflect the fact that problem descriptions are often dynamic, evolving in the course of abductive problem solving. The problem description includes or excludes observations on the basis of their relevance to some categorization of the event or some commitment to a range of potential explanatory mechanisms. If, as often happens, this categorization or commitment shifts in the course of inquiry, so also does the problem description.

Peirce applies the concept of abduction not only to judgment but also to perception. Perceptual abduction is necessarily type or pattern-based:

> Abductive inference shades into perceptual judgment without any sharp line of demarcation between them.... The abductive suggestion comes to us like a flash. It is an act of insight, although of extremely fallible insight.
>
> [The form of the perceptual abduction is]:
>
> A well-recognized kind of object M, has for its ordinary predicates P1, P2, P3,..., etc., indistinctly recognized.
>
> The suggesting object S has these same predicates, Pl, P2, P3, etc. Hence S is of the kind M. (Peirce, 1955, p. 305)

The morphological or type-based explanation described here poses certain difficulties for logic-based models. The problem is that sometimes types only loosely determine the characteristics of the typed object. For example, I see a strange looking red fruit on my neighbor's tree and after a moment's reflection hypothesize that the fruit is an unfamiliar variety of apple. In this case the application of the type *apple* does not entail the observed redness (since apples can also be yellow or green) yet the redness contributed to my identification of the fruit as an apple. Thus the process of discovering the type or pattern which explains the object is not well modeled by reasoning backward through a rule. It is something more like the extrapolation of a pattern or type from a partial set of features.

In the passage quoted at the beginning of section 4, Peirce mentions the "surprising" character of the facts which trigger abductive perception or judgment. This refers to the fact that abduction arises only when a set of facts is not already explained or when the type or pattern to which an object belongs is not already apparent. The 'surprise' is based on a background of familiar patterns, relationships, and expectations. For example, when the leaves on my apple tree acquire a yellow tint in mid-summer rather than mid-autumn I take that as something requiring explanation, a potential indicator of an unhealthy condition. When the needle on the temperature gauge in my car remains in a fixed position this indicates a potential problem. Based on the interplay of observations and expectations certain facts become significant either as explanatory goals or clues in an investigation. Logic-based abduction does not account for the important role of abnormality, absence, and

4. THE PEIRCEAN ACCOUNT OF ABDUCTION

The modem concept of abduction is frequently traced to the philosopher C.S. Peirce. According to Peirce abduction is:

> the operation of adopting an explanatory hypothesis [where] the hypothesis cannot be admitted even as a hypothesis, unless it be supposed that it would account for the facts or some of them. The form of inference is therefore this.
>
> The surprising fact C, is observed; But if A were true, C would be a matter of course, Hence there is reason to suspect that A is true. (Peirce, 1955, p. 151)

This often cited passage leaves a number of important questions unanswered: What is the relation between an 'explanatory hypothesis' and an actual explanation? In what sense should the facts to be explained follow as 'a matter of course' from the explanatory hypothesis? By what criteria ought the quality of an explanation be judged? What is the nature of the epistemic warrant supplied by an abductive 'inference'? Researchers in AI have tended to avoid these 'philosophical' questions in favor of a more practical goal: providing a model for the construction of explanatory hypotheses. We have seen, however, that the failure to attend more closely to certain philosophical issues may have cut this work off from some potential sources of insight.

The first and most crucial issue that needs to be addressed is the precise sense in which explanatory hypotheses must 'account for the facts'. In terms of the framework laid out in section 2, what types of covering laws are required by abductive explanation? There has been a tendency in AI circles to reduce abductive explanation to causal explanation. Peirce clearly does not follow this course.

Consider Peirce's examples. In one case he describes how observed qualities of dress, bearing, and gait led to a hypothesis that the man before him was a priest. He also describes how Kepler extrapolated an orbit for Mars. Guided by Copernican theory, Kepler postulated an elliptical shape to account for a collection of observations of the position of Mars at different moments in time (Peirce, 1955, p. 152-155). At another point in his writings (Peirce, 1958, p. 479) Peirce speaks of landing at a Turkish seaport and seeing a man on horseback surrounded by four horsemen holding a canopy over his head. From the remote location, the great display of honor shown to the central figure, and the unlikelihood of finding exogenous personages of great rank at such a location, Peirce correctly hypothesized that the central figure must be the governor of the province. In each of these cases what is abduced is 1) an object type (priest, provincial governor) whose properties would account for a set of observations or 2) a pattern or rule (elliptical heliocentric orbit) which would account for a set of observations. It would clearly be stretching things to assimilate these examples to the abduction of causes. If causal explanation were to be invoked here at all it would only be at a secondary stage to explain the source of the general pattern or type.

When Peirce writes that an explanatory hypothesis "would account for the facts or some of them" he acknowledges that successful explanation may indeed only account for some of the facts. The adoption of an explanatory hypothesis selects or circumscribes those facts about a situation that are significant. For example, suppose the man later identified as a priest was observed wearing a pair of eyeglasses. The observer will probably not be able to explain the eyeglasses from the priest hypothesis. Only after the fact, however, does it become clear

violated expectation in focusing abductive search. Consistency-based approaches are stronger in this area (Poole, 1989). We believe an adequate account of abductive problem solving must illuminate its expectation-driven character.

5. HIERARCHICAL ABSTRACTION IN ABDUCTIVE PROBLEM SOLVING

Cognitive models in AI frequently assume a given problem description. Problem solving in the model proceeds relative to the vocabulary of that description. This ignores a number of issues affecting human problem solving. Human problem solvers often begin in a poorly defined situation. They must first choose a descriptive vocabulary reflecting their preliminary understanding of the character of the situation and the type of problem solving methods appropriate. They must also select a level or description, or equivalently, a level of abstraction, determining what sorts of detail are currently appropriate. Problem solving then proceeds interactively, with this initial description refined and revised based on information gathered along the way. The process of description construction is generally very knowledge intensive, depending as heavily on domain knowledge and skill as the reasoning based on that description.

We introduce the notion of hierarchically structured investigation by glancing at two somewhat familiar examples of investigative processes:

1) Investigators of an airplane crash attempt to determine as soon as possible whether the crash was a result of mechanical failure, bomb, fire, pilot error, or mid-air collision. The focus initially is the *type* of causal mechanism, and details are considered primarily to the degree to which they have bearing on this issue. Once the type of causal mechanism is identified, the description of relevant facts and the focus of inquiry shift. If the cause of failure has been determined to be mechanical failure, for example, certain facts which earlier were deemed to be potentially significant (e.g., the presence of other planes in the vicinity), recede into the background, while others, such as structural design and maintenance history, come under closer scrutiny.

2) Police investigating a death also try to determining quickly the *type* of operative causality: was the death an accident, a homicide, a suicide, etc.? If it has been determined that the cause of death was homicide, the focus of inquiry shifts. The person's medical history becomes less significant than evidence such as fingerprints, forced windows, and family or business relationships. Identifying the type of causal mechanism determines the things that need to be discovered in order to complete the explanation. If it is a murder, then murderer, motive, time, place, weapon, and method represent 'slots' whose values need to be determined.

In our recent work with experts in discrete semiconductor failure analysis (Luger & Stern, 1992), we have encountered a well defined but complex problem solving structure similarly based on a hierarchy of abstractions. At the center of this structure is a set of explanation schemata or templates which organize the problem solving process. In the remainder of this section, we will attempt to describe this organization and structure.

Sandia semiconductor engineers are required to perform failure analysis when there have been a significant number of failures involving the same device. The devices are used in 'critical' applications, so their failures must be carefully monitored to minimize the possibility of eventual field failures.

What contributes to the difficulty of failure analysis is that device failures have widely varying types of explanation. This condition arises partly because failures which come to failure analysis can occur in different phases of the devices's life cycle. The failures can occur at the manufacturer, during screening and acceptance testing, after the device is in a circuit assembly, as a latent field failure, etc. This means that explanations can include factors as diverse as: poor materials quality, faulty fabrication process, bad device design, bad circuit design, inadequate screening, destructive handling, contamination, electrostatic discharge, electrical overstress, improper storage, etc.

To master the complexity of this search space, the problem solving process is partitioned into phases. The initial stages are concerned with gathering evidence from which a high level characterization of the causal mechanism can be constructed. Once the type of causal mechanism is characterized at a high level, a new phase is entered which attempts to validate and refine the earlier characterization. If the new phase is successful the earlier characterization is replaced by terms belonging to a lower level of description, and the cycle is repeated.

During this problem solving process the description of the failure situation evolves dramatically as the investigation progresses. This involves more than merely the acquisition of new data. Facts which were earlier included in the situation description are no longer mentioned. Observed but unregistered features of the situation (e.g., location, duration, degree of a quality) are now explicitly formalized because of their potential significance in the current problem solving environment. The same facts which were earlier expressed in general terms are now given a finer grained description, recategorized to match the slots of a new explanation schema.

A correlation can be observed between the forms of explanation and the practical goals of the problem solver. The high level goal in discrete semiconductor failure analysis is not to remediate past failures; these are corrected when necessary by simply replacing the failed part with a good one. The goal is rather to understand the cause of failure well enough so that a high degree of quality control can be assured. Different failure mechanisms generally have different methods for restoring quality control ("cures"). Tester induced overstress is corrected by identifying and repairing the faulty test equipment or by changing test procedures. Atmospheric contamination from package leaks can sometimes be handled by imposing a more rigorous hermetic seal screen and sometimes by changing to a different package. Not so curiously, the explanation schema associated with each type of failure mechanism generally possesses precisely that structure and level of detail required to guide the associated "cure".

We have also observed a correlation between the quality of problem descriptions and the problem solver's domain knowledge and problem solving skills. The expert's descriptions are generally more selective, pertinent and coherent than that of less experienced engineers. The expert's interpretive mechanisms are clearer and more highly structured than that of the novice. Through the years these experts have had repeated experience with a wide range of failure mechanisms. They appear to have abstracted from this experience an ability to recognize *types* of failure mechanisms from patterns of evidence. We attribute this ability to the possession of a richer set of *explanation templates* than that possessed by inexperienced engineers. It is the possession of such explanation templates that allows successful problem solvers to recognize from the outset which aspects of a situation are potentially significant.

In the initial phase of problem solving, explanatory goals are very abstract and schematic. The problem with the semiconductor may be due to electrical overstress, electrostatic discharge, a manufacturing defect, a particle induced short, internal volatile contamination, etc. First the initial description is used to prune the set of goals; then on the basis of the remaining goals, data gathering begins. Observation and data gathering is goal driven: observations are made which will validate or invalidate current hypotheses. These same observations also serve to define and articulate the hypotheses, providing values for the slot descriptors. Thus the explanatory goals serve as a basis for reasoning which generates tests and organizes the investigation.

At each stage of the investigation, the results of data gathering are folded back into the problem description, further defining the material which a satisfactory hypothesis must explain. At certain stages of the investigation, when sufficient strength attaches to a particular hypothesis, the entire process cycles, moving to a deeper and more detailed level of description. In transistor failure analysis, for example, reasoning initially revolves around the external electrical and mechanical characteristics of the failed device. Later, however, when a well supported hypothesis is in placé, the device can be opened and examined under a scanning electron microscope. At this point, the expert has a better idea of where to look and what to look for. This inspection may yield a precise image of the damage on a junction, with its location and geometry, or a precise identification of the composition of some foreign particle.

Suppose, for example, the electrical symptom is high reverse leakage. If other causes have been eliminated, and if the context indicates potential exposure to electrostatic discharge (ESD), and if the device is ESD vulnerable, then ESD damage might become the strongest hypothesis. Suppose now the transistor package is opened and the device is examined under a scanning electron microscope. The purpose is both to confirm the hypothesis of ESD damage and also to more fully characterize the nature of the damage. Perhaps traces of an arc in the oxide layer are observed. This arc trace represents a damaged crystalline structure. The hypothesis of ESD damage is thus corroborated. At the same time explanation has been reconstructed at a finer level of detail, one closer to the first principles of the domain. This finer description, including the location and extent of ESD damage, can finally be used to identify or constrain the environmental source of electrostatic discharge which caused the damage.

This example illustrates an important kind of explanation expansion or deepening. In the logic-based approach, explanations are deepened by backchaining on causal conditions, i.e. by finding causes of causes. In this example, however, a different kind of deepening occurs. We start from the symptom of high reverse leakage, explaining it using the rule ESD damage causes high reverse leakage'. This rule represents at best a statistical regularity, a regularity that itself needs to be explained. In the logic-based approach the only way to deepen this explanation is by finding a cause for ESD damage. This explains at best the origination of the cause, not the nature of the causal mechanism by which the cause produced the effect. We see in the example above, however, how newly acquired information allows for the articulation of a more detailed and intelligible set of relationships: 1) ESD from a human source caused an arc along the emitter-base oxide layer; 2) This arc caused a .01 micron leakage path for electrons; 3) Electron flow along this leakage channel causes a measurable reverse leakage. Moving to a more detailed level of description here provides more intermediate links between a more fully characterized cause and effect.

This example also illustrates another feature of abductive problem solving. Both the situation description and the gathering of evidence are informed by explanatory hypotheses. As the explanatory focus narrows and deepens, the descriptive form comes to more and more closely approximate the form of a particular explanation schema. At the end of this process, if it is successful, a partial convergence of hypothesis and description occurs, so that the description itself comes to play the role earlier played by the explanatory hypothesis.

6. TOWARDS A COMPUTATIONAL MODEL OF ABDUCTIVE PROBLEM SOLVING

We have seen how description, hypothesis construction, and data gathering alternate. These are in turn tied to practical problem solving efforts. The results of unsuccessful or partially successful attempts at problem intervention become evidence which can, in turn, serve to reject or refine explanatory hypotheses. We suggest that abductive problem solving must thus be viewed as an integrated process. The logic of abductive problem solving is a "logic" of interrogation and discovery. It is organized, we argue, by "explanation templates". We now give a brief account of how we have attempted to model this organizing role of explanation templates in abductive problem solving.

The expert system we are building initially guides the user through a series of electrical tests of the failed device. The purpose of this is both to confirm the reported failure mode and also to provide an evidential basis for invoking a set of abstract top-level explanation templates. The electrical characteristics of the failed device are compared with those of a normal device and particular abnormal electrical characteristics are used as signs or indicators of potential failure mechanisms. For example, high reverse leakage is taken as a potential indicator of ESD damage, volatile contamination, or electrical overstress.

Each failure mechanism has an abstract explanation template associated with it. Explanation templates consist of:

1) A set of entry conditions, defining the kinds of situations to which an explanation template is applicable;
2) A set of slots specifying the elements of the explanation, with rules defining allowable values for each slot;
3) A set of validation procedures for confirming or disconfirming the hypothesis, delineating tests or observations based on the expected consequences of the hypothesis;
4) Links to other, lower level, explanation schemata which expand part or all of the explanatory hypothesis.

Once the initial set of explanation templates has been activated, pruning begins. The activation conditions of a template must be satisfied. For example, for ESD damage to be a viable hypothesis, the device must be ESD sensitive, it must have failed in an environment where a sufficiently strong ESD source was present, and it must have been unprotected. If any of these conditions is known to be false, the hypothesis can be discarded.

The remaining hypotheses then organize the next phase of evidence gathering. Each template has a set of rules for building an explanation by filling in its slots. The discovery of incompatible slot values causes the explanation to be discarded. Once slot values have been determined, the hypothesis validation stage begins. The presence of a causal condition

or mechanism usually has predictable consequences. These consequences provide the basis for further investigation and testing. For example, if volatile contamination were responsible for high leakage, then certain electrical characteristics of the device would be temperature sensitive. Thus a natural test for the hypothesis of volatile contamination involves baking the failed device at high temperatures and then retesting to see if that changes or improves a degraded characteristic. Observe that hypothesis validation often involves forward reasoning from an hypothesized condition to effects which would follow if that condition were present. The interweaving of forward and backward reasoning in abductive search is a characteristic that is overlooked by most logic-based models.

Once the set of explanatory hypotheses has been pruned and the remaining hypotheses have been corroborated to whatever extent possible, each explanation is expanded and deepened by chaining to a *child* template. This child template generally has a larger set of slots, allowing for a finer grained, more detailed explanation. The slot values of this child template must be determined, either by inferring them from that of the parent using rules for redescription, or by further observation and investigation. For example, the expanded form of the volatile contamination template contains slots for contaminant composition, source, concentration, and degrading mechanism (the mechanism by which the contamination produces a degradation of the device's electrical characteristics). In this case, there are not one but two child templates, one for corrosive and one for non-corrosive contaminants. Which descendent template is invoked is decided either by determining the contaminant composition or by discovering signs of corrosion.

Finally, explanation templates at the lowest level are linked with "cures". Slot values at this level appear as antecedent conditions in rules that determe how to resolve a failure syndrome or restore quality control. The failure or rejection of a recommended cure can trigger reevaluation of the explanation upon which that cure was based.

7. CONCLUSIONS AND FUTURE RESEARCH

As the Sandia example illustrates, investigative problem solving is often highly interactive in nature. Data gathering and hypothesis construction are tightly interwoven, with each determining the course of the other. Explanation templates provide a data structure which supports this interactive process. They facilitate cost effective evidence gathering and efficient search.

Logic-based models of abduction, unfortunately, employ a fixed description of the problem solving situation. The absence of pertinent information in this initial description drastically increases the complexity of search by contributing additional degrees of freedom to the parameters of explanation. This complexity problem is handled in the template-based model by abstraction. Search for an explanation only proceeds at a single level of description at a time using evidence to decide between the currently competing explanation schemas or subschemas. This allows for lazy evaluation: tests or decisions regarding the values of subschema parameters need not be made at the early stages of an investigation but can be postponed until later.

In our ongoing research we are attempting to solidify the explanation template model of abduction. The problems we see which need to be addressed are:

1) How can explanation templates be characterized at the *knowledge level*, independently of particular data structures? How can they be formalized? Is there a *logic* of explanation?

2) How are explanation templates acquired? What is the abstraction mechanism which allows new explanation templates to be abstracted from novel cases?
3) How can explanation templates be combined to construct compound or hybrid explanations?
4) How can we better model the complex pattern recognition capabilities that humans use in invoking explanation templates?
5) When and how can explanation templates be derived from *explanation from first principles*?

Acknowledgments: This research was partly supported by a grant from Sandia National Laboratories. The Discrete Failure Analysis Expert System has been developed in conjunction with SNL. We wish to give special thanks to the Sandia semiconductor experts, Allan E. Asselmeier and Don Holck.

NOTES

1. Sound reasoning is reasoning which is guaranteed to yield true conclusions from true premises. Unsoundness, however, is not necessarily a fatal flaw. Unsound modes of inference can often be both useful and appropriate.
2. Logic-based models often include an 'explanation selection' phase, during which an attempt is made to sift out the best explanation(s). But a) this generally requires use of extra-logical criteria, and b) this generally occurs only after the cost of generating unsuitable explanations has already been paid.
3. *ATMS* stands for *assumption based truth maintenance system.* ATMSs can be used to perform diagnostic reasoning and other abductive tasks. See de Kleer (1986).
4. Observe that incomplete explanations, explanations in which the *explanans* does not logically entail the *explanandum*, can be restated as complete explanations with incomplete covering laws, that is, covering laws which are false as stated because the application conditions are left unspecified.

REFERENCES

Charniak, E. & Shimony, S. 1990. Probabilistic Semantics for Cost Based Abduction. *Proceedings of the Eighth National Conference on Artificial Intelligence,* (pp. 106–111). Menlo Park CA: AAAI Press / MIT Press.

de Kleer, J. 1986. An assumption-based TMS, *Artificial Intelligence,* 28:127–162, Amsterdam: North Holland.

Doyle, J. 1983. Methodological Simplicity in Expert System Construction: The Case of Judgments and Reasoned Assumptions. In Shafer & Pearl (Eds.), *Uncertain Reasoning.* 1990. San Mateo: Morgan Kaufman, pp. 689–693.

Eco, Umberto. 1976. A *Theory of Semiotics.* Bloomington, IN: Indiana University Press.

Gallanti, M., Roncato, M., Stefanini, A., Tornielli, G. A Diagnostic Algorithm based on Models at Different Levels of Abstraction. *Proceeding of the Eleventh International Joint Conference on Artificial Intelligence,* (pp. 1350–1355). San Mateo, CA.: Morgan Kaufman.

Hempel, C.G. 1965. Studies in the logic of confirmation. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science.* New York: The Free Press, pp. 3–51.

Israel, D.J. What's wrong with non-monotonic logic? 1980. *Proceedings of the First Annual National Conference on Artificial Intelligence.* Los Altos, CA: Morgan Kaufman.

Josephson, J. R., Chandrasekaran, B., Smith, J. W. Jr, & Tanner, M. C. (1987). A Mechanism for Forming Composite Explanatory Hypotheses. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC *17*(3), 445–454.

Levesque, H. J. (1989). A knowledge level account of abduction. *Proceeding of the Eleventh International Joint Conference on Artificial Intelligence*, (pp. 1061–1067). San Mateo, CA: Morgan Kaufman.

Luger, G. F. & Stern, C. (1992). Expert Systems and the abductive circle. In R. Jorna & B. von Heudsen (Eds.) *Cognition and Semiotics*. Berlin: Walter de Gruyter.

Ng, H. T. & Mooney, R. J. (1990). On the Role of Coherence in Abductive Explanation. *Proceedings of the Eighth National Conference on Artificial Intelligence*, (pp. 337–342). Menlo Park CA: AAAI Press / MIT Press.

Pearl, J. (1987). On Evidential Reasoning in a Hierarchy of Hypotheses. *Artificial Intelligence* 28 (1986) 9–15. Reprinted in I. Shafer & J. Pearl (Eds.), *Uncertain Reasoning*. 1990. San Mateo: Morgan Kaufman, pp. 449–451.

Peirce, C.S. (1955). *The Philosophical Writing of Peirce*. New York: Dover Publications.

Peirce, C.S. (1958). *Collected Papers* 1931–1958. Cambridge: Harvard University Press.

Peng, Y. & Reggia, J. A. (1986). Plausibility of Diagnostic Hypotheses: the Nature of Simplicity. *Proceedings of the Fifth National Conference on Artificial Intelligence*, (pp. 140–145). Philadelphia, PA.

Poole, D. (1989). Normality and Faults in Logic-Based Diagnosis. *Proceeding of the Eleventh International Joint Conference on Artificial Intelligence*, (pp. 1304–1310). San Mateo, CA: Morgan Kaufman.

Reggia, J., Nau, D. S., Wang, P. Y. (1983). Diagnostic expert systems based on a set covering model. *International Journal of Man-Machine Studies. 19*(5), 437–460.

Reiter, R. (1987). Nonmonotonic Reasoning. *Annual Revue of Computer Science*. Reprinted in I. Shafer & J. Pearl (Eds.), *Uncertain Reasoning*. 1990. San Mateo: Morgan Kaufman, pp. 637–656.

Reiter, R. & deKleer, J. (1980). Foundations of assumption-based truth maintenance systems. *Proceedings AAI-87*, Seattle, WA, 1987, pp. 183–188.

Selman, B. & Levesque, H. J. (1990). Abductive and Default Reasoning: A Computational Core. *Proceedings of the Eighth National Conference on Artificial Intelligence*, (pp. 343–348). Menlo Park CA: AAAI Press / MIT Press.

Thagard, P. 1989. Explanatory Coherence. *Behavioral And Brain Sciences 12*, 435–502.

Address all correspondence to: Carl Stern, Department of Computer Science, University of New Mexico.