Abduction and Learning: Case Studies in Diverse Domains

Paul O'Rorke
ororke@ics.uci.edu

Technical Report 92-98

September 21, 1992

Research supported in part by National Science Foundation Grant Number IRI-8813048, by Douglas Aircraft Company and the University of California Microelectronics Innovation and Computer Research Opportunities Program, and by an Irvine Faculty Fellowship from the University of California, Irvine Academic Senate Committee on Research.

Abduction and Learning: Case Studies in Diverse Domains

Paul O'Rorke[†]

Phone and Fax: (714) 854-2894
Electronic Mail: ororke@ics.uci.edu
Department of Information and Computer Science
University of California, Irvine, CA 92717
United States of America

September 21, 1992

[†]Research supported in part by National Science Foundation Grant Number IRI-8813048, by Douglas Aircraft Company and the University of California Microelectronics Innovation and Computer Research Opportunities Program, and by an Irvine Faculty Fellowship from the University of California, Irvine Academic Senate Committee on Research.

Abstract

This paper presents a knowledge-based learning method and reports on case studies in different domains. The method integrates abduction and learning. Abduction provides an improved method for constructing explanations, in comparison with deductive methods traditionally associated with explanation-based learning. The improvement enlarges the set of examples that can be explained so that one can learn from additional examples using explanation-based macro-learning techniques. Abduction also provides a form of knowledge level learning. The importance of abductive learning is shown by case studies involving over a hundred examples taken from diverse domains requiring logical, physical, and psychological knowledge and reasoning. The case studies are relevant to a wide range of practical tasks including: natural language understanding and plan recognition; qualitative physical reasoning and postdiction; diagnosis and signal interpretation; and decision-making under uncertainty. The descriptions of the case studies show how to set up abduction engines in particular domains and how abduction solves particular tasks. They show how to provide and how to represent the relevant knowledge, including both domain knowledge and meta-knowledge relevant to abduction and search control. The description of each case study includes an example, its explanation, and discussions of what is learned by macro-learning and by abductive inference.

KEYWORDS: *abduction, explanation-based learning*

Contents

1 Introduction **1**
- 1.1 Inference and Learning 1
- 1.2 Abductive Inference 1

2 Abduction and Learning **2**
- 2.1 Abductive Macro Learning 2
- 2.2 An Example: Liver Diagnosis 6

3 Case Studies **9**
- 3.1 Case Study: Explaining Emotions 10
- 3.2 Case Study: Explaining Physical Processes 16
- 3.3 Case Study: Explaining Decisions 21
- 3.4 Case Study: Explaining Signals 25
- 3.5 Summary of Case Studies 26

4 Discussion **29**
- 4.1 Abductive Hypothesis Formation 29
- 4.2 Assumability and Operationality 31
- 4.3 Search Control 32
- 4.4 Macro Learning 33
- 4.5 The Value of Learning 35

5 Related Work, Limitations, and Future Work **35**

6 Conclusion **37**

List of Figures

1 Inputs and Outputs of the Abduction Engine 3
2 An Explanation for Pople's Liver Diagnosis Example 8
3 An NDE Test and the Resulting Signal 25

List of Tables

1 A Procedural Sketch of the Abduction Engine 4
2 Background Knowledge for Pople's Liver Diagnosis Example 6
3 A Partial Explanation for Pople's Liver Diagnosis Example 7
4 A Macro Learned from Pople's Liver Diagnosis Example 9
5 A Special Case of the Learned Macro 9
6 Elicitation Conditions for 20 Emotion Types 11
7 Explanations for Relief and Fear 13
8 Laws of Qualitative Process Theory 17
9 A Qualitative Theory of Combustion and Calcination 17
10 Explanation of an Increase in the Weight of Mercurius Calcinatus 19
11 Macro Learned in the Case of Mercurius Calcinatus 20
12 Some Rules of Real Arithmetic 22
13 The B-1B Example 23
14 Explanation of the Expert's Recommended Action 24
15 Explanation of an Anomalous NDE Signal 27
16 Macro Learned from the Interpretation of the NDE Signal 28

1 Introduction

This paper presents an abductive form of explanation-based learning. In this introductory section, we provide some background information on abduction and on the role of inference in learning. In the next section, we provide a description of our learning method and an illustration in terms of a concrete example. Our goal in doing this is to provide a description sufficiently detailed to enable readers to implement and use the method. In the subsequent section, we describe case studies applying this method in four different domains. One goal of the case studies section is to show how the method has been applied in sufficient detail to enable the reader to apply the method in new domains. Another goal is to draw general lessons from the case studies. Next, we discuss related work, limitations of the present work, and suggestions for future work. The final section summarizes the paper and draws conclusions.

1.1 Inference and Learning

Most machine learning methods involve some form of *induction* or inference from specific to general statements; consequently they are often called "data-driven," "empirical," or "similarity-based" learning methods (see, e.g., Michalski & Chilausky, 1980; Quinlan, 1986). Recently, attention has been given to a complementary class of "knowledge-driven," "analytical," or "explanation-based" (EBL) learning methods (see e.g., DeJong & Mooney, 1986; Mitchell, Keller, & Kedar-Cabelli, 1986) but these methods have been characterized in terms of *deduction.* There is a third form of inference called *abduction.* We argue in this paper that abductive inference is at least as fundamental and important for learning as inductive and deductive inference.

Intuitively, EBL is "learning based upon explanations," so EBL theories and systems include components aimed at describing or implementing processes for constructing explanations. Formalizations and implementations of EBL can be improved upon by introducing more sophisticated models of the explanation process. We argue that an important step in this direction is to view explanation as a kind of plausible inference process — *one that is not often deductive.* The particular form of plausible explanatory inference explored here is based upon Peirce's notion of abduction.

1.2 Abductive Inference

Charles Sanders Peirce (Peirce, 1931–1958) used the term abduction as a name for a particular form of explanatory hypothesis generation. His description was basically:

> *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.*

In other words, if there is a causal or logical reason A for C, and C is observed, then one might conjecture that A is true in an effort to explain C.

Since Peirce's original formulation, many variants of this form of reasoning have also come to be referred to as abduction. We focus on a logical view of abduction advocated by Poole (e.g., Poole, Goebel, & Aleliunas, 1987).[1] In this approach, observations O are explained given some background knowledge expressed as a logical theory T by finding some hypotheses H such that

$$H \wedge T \vdash O.$$

In other words, if the hypotheses are assumed, the observation follows by way of general laws and other facts given as background knowledge. Consistency is also desirable so it is usually required that

$$H \wedge T \nvdash \mathit{false}.$$

2 Abduction and Learning

In the influential model of EBL presented by Mitchell et al (1986) and in implementations such as (Kedar-Cabelli & McCarty, 1987), learning is based upon explanations generated by a deductive theorem prover. The learning method is essentially a form of lemma caching or deductive macro-learning.

This form of EBL has been criticized on the grounds that it only improves efficiency and does not involve "learning at the knowledge-level" (as defined by Dietterich, 1986). The deductive closure of the knowledge-base does not change as a result of learning because the macro-learning method specializes existing general knowledge, even though it generalizes given examples.

This early model of EBL rests on a purely deductive model of explanation. Integrating more sophisticated models of the explanation process with learning leads to interesting new models of EBL with additional learning capabilities.

2.1 Abductive Macro Learning

We present an integration of abduction and learning that combines a first order logical form of abduction with macro-learning. The abduction component is based on an early approach to mechanizing abduction described in (Pople, 1973). The method is implemented in a PROLOG meta-interpreter called AMAL (Abductive MAcro Learner).[2]

[1] See also (Levesque, 1989).

[2] PROLOG was chosen because the basic operation involved in constructing explanations, abductive inference, is similar to backward chaining. PROLOG provides basic operations such as unification that are an essential part of backward chaining.

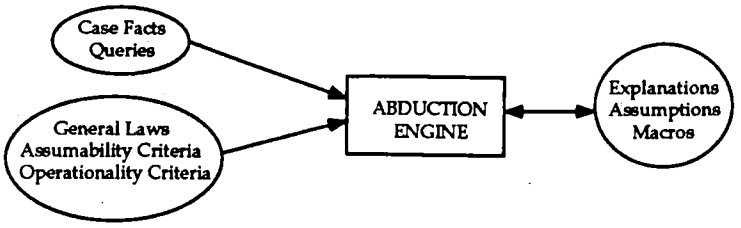

Figure 1: Inputs and Outputs of the Abduction Engine

An input/output characterization of the program is given in Figure 1. AMAL takes as input a collection of PROLOG clauses encoding theories. One theory represents background knowledge, another captures the facts of the case at hand. An observation to be explained is given as a query.

AMAL is also given an operationality criterion and an assumability criterion. The operationality criterion is used to flag queries that should be turned over to the underlying PROLOG interpreter. The intuition is that AMAL performs explanatory reasoning, whereas the PROLOG interpreter performs lower-level reasoning in a more efficient manner, without keeping track of explanatory relationships. Separate theories are provided: explanatory clauses are used to construct explanations, while ordinary PROLOG clauses are used for operational computations.

The assumability criterion determines whether a query that could not be proved or disproved may be assumed. The query may or may not be operational.[3] In general, the assumability criterion is used to reject inadmissible hypotheses and to assume admissible hypotheses. The user provides characterizations of admissible and inadmissible hypothesis as part of the domain description but in some domains in questionable cases the user is asked whether a specific hypothesis should be assumed.

AMAL's output includes an explanation of the given observation. The explanation can include assumptions made in order to complete the explanation. These assumptions, plus a macro learned by traditional explanation-based macro learning, are added to the knowledge base for use on subsequent examples.

A description of the procedure followed by the abduction engine is shown in Table 1. The abduction engine attempts to construct explanations of given observations using general laws and specific facts. In the implementation, explanations are proofs represented as AND trees. Observations to be explained correspond to conclusions of the proofs (roots of the trees). General laws are encoded as rules and these are used to generate the proofs through a process based upon backward chaining.

The mechanization of abduction is comprised of three steps (Table 1). The first step

[3]In most cases, assumptions involving operational hypotheses are disallowed. However in some domains they are allowed (see Section 3.1).

Table 1: A Procedural Sketch of the Abduction Engine

Given: a first-in-first-out queue of queries Q containing observations to be explained;
Find: explanations for the queries, using abductive inferences.

1. BACKWARD CHAINING:

   While the query list Q is not empty, do:

   (a) Select the first query q and remove it from Q.
   (b) If q is operational then compute an answer directly if possible, else
   (c) If q is an admissible goal and is indirectly explainable using a rule,
   then use the rule to generate new queries and add them to Q, else
   If q is an admissible hypothesis
   then add q to the list U of unexplained queries,
   else fail and backtrack.

2. IDENTIFICATION:

   While there are unifiable pairs of queries in U, unify and replace pairs.

3. ASSUMPTION:

   While there are unexplained queries in U,

   (a) Select the first query u and remove it from U.
   (b) If the truth of u is not known, and it is an admissible hypothesis, and it is ratified by the user (optional), then assume u, else fail and backtrack.

corresponds to backward chaining as it is implemented in PROLOG interpreters. The observation is treated as a query. Initially, there is only one query but in general, there may be a number of open questions in the query list Q. The search process attempts to ground the explanation tree in known facts. If a query is operational, AMAL attempts to identify it with a fact in the data-base or in its deductive closure. In attempting to prove operational queries, AMAL does not keep track of an explanation and it does not use "explanatory" clauses. However, it does allow for the possibility that a query may be operational and/or provable in several ways. If one operationalization of the query fails to pan out, backtracking is used to search for another.[4] If the query is not operational, or no direct operational explanation is possible, then explanatory rules may be used to extend the partial explanation, replacing existing queries with new queries. Before queries are allowed to generate new queries in this manner, a domain-dependent test is applied so that goals deemed inadmissible are disallowed. This is an important form of search control (see Section 3.3). A similar domain-dependent test is applied to reject inadmissible hypotheses among the remaining queries.

The second step begins when backward chaining fails. In this step, the remaining unexplained queries are examined and some of them are assumed to be "the same." Unlike the previous step, this inference is not deductively sound, but errors are recoverable through backtracking. In terms of Table 1, at the beginning of this stage Q is the empty list, U is a non-nil list of unexplained statements, and the explanation is incomplete. The algorithm continues by first selecting an arbitrary unexplained statement u from U. If u can be identified (unified) with any other statement in U, then the pair is replaced in U with their identification. The identification step ends when no more queries in U are pairwise identifiable.

This "identification" or "merging" step is based on the *synthesis* operator advocated by Pople (1973) and justified in terms of Occam's razor. This operation simplifies explanations, reduces the number of assumptions that must be made, and increases the plausibility of the explanation. Another advantage is that identification assumptions often introduce new information. The identification of two previously unrelated statements in different parts of an explanation often causes sharing of information between the separate branches of the explanation. In the implementation, statements are identified by unifying two well-formed-formulae. This can cause variables in both formulae to take on new bindings. The new bindings then propagate to other statements that share the same variables. See Section 2.2 for an example of this sort of information sharing.

The third abduction step tests whether remaining queries can be assumed. The queries are tested to ensure that they are not known to be true or false. Non-explanatory theorem

[4]The notion of operationality used here is relatively flexible; it takes advantage of an underlying theorem prover capable of reasoning about operationality. See Hirsh (1987) for a discussion of the importance of this feature.

Table 2: Background Knowledge for Pople's Liver Diagnosis Example

inflammatory(abscess).
located_in(liver,right_upper_quadrant).
chills ← *present(P,S), inflammatory(P).*
pain(R) ← *present(P,S), located_in(S,R).*
poples_syndrome(R) ← *chills, pain(R).*

proving is allowed in testing whether a hypothesis is known to be true. AMAL calls PROLOG and if the hypothesis is proven true then it is not allowed as an assumption. A test against stored negative assertions is used to determine whether a hypothesis is false (we do not use negation as failure). This test is a limited form of the consistency check called for in the formal specification of abduction (see Section 1.2). Together, these two tests ensure that a hypothesis is not known to be true or false. Next, an "assumability" test is used to decide whether to assume that a hypothesis is true. The test includes a domain-independent component and a hook that takes advantage of domain-dependent information about admissible hypotheses. A human user may also be consulted in some domains. This test is applied to each of the queries u in list U. If u is not assumable, then the current attempt to find an explanation is aborted and backtracking is invoked in order to continue the search for acceptable explanations.

2.2 An Example: Liver Diagnosis

An example adapted from Pople (1973) serves to illustrate the method. Suppose the task is to do diagnosis by explaining observed symptoms in terms of underlying disease(s). Suppose that the general statements shown in Table 2 are given. The rules state that chills may be caused by the presence of an inflammation in some region of the body. Furthermore, pain in a region may be caused by the presence of a problem in a structure in that region. The facts state that the liver is in the right upper quadrant and abscesses are inflammatory.[5]

Given an observation encoding the symptoms *chills* and *pain* in the right upper quadrant, the abduction system attempts to explain these symptoms by backward chaining on the given rules, attempting to ground out in known facts. Table 3 shows a partial explanation. At this point, the chills are explained as the result of an inflammatory abscess while

[5]Some of these facts (e.g., that abscesses are inflammatory) might be better stated as universally quantified implications (e.g., $\forall X, inflammatory(X) \leftarrow abscess(X)$). We follow the original formulation in order to keep the example as simple as possible.

Table 3: A Partial Explanation for Pople's Liver Diagnosis Example

poples_syndrome(right_upper_quadrant)
chills
present(abscess, _73)
inflammatory(abscess)
pain(right_upper_quadrant)
present(_232, liver)
located_in(liver, right_upper_quadrant)

the pain in the right upper quadrant is explained by the presence of a problem in the liver. The fact that abscesses are inflammatory contributes to the explanation of chills while the fact that the liver is in the right upper quadrant contributes to the explanation of pain in the right upper quadrant. This explanation is incomplete: Two questions remain — *is something present in the liver?* and *is there an inflammatory abscess present somewhere in the body?* These questions cannot be answered using only the given facts. This is indicated by the presence of boxes and variables in the table. The boxes indicate queries, open questions, or gaps in the explanation. They identify hypotheses that, if assumed, would complete the explanation. Variables in the explanation correspond to unidentified objects. They are existentially quantified. In this case, the hypotheses state that there is something in the liver and there is an inflammatory abscess somewhere in the body.

An explanation-based learning system based exclusively on deduction (Hirsh, 1987; Kedar-Cabelli & McCarty, 1987; Prieditis & Mostow, 1987) will fail to explain, generalize, and learn from this example because there is no proof of chills and pain from the known facts. AMAL does *not* fail. Instead, it completes the explanation by going beyond deduction in two ways: (1) it identifies the presence of an abscess in some bodily structure with the presence of something in the liver; and (2) it assumes that there is indeed an inflammatory abscess in the liver. The result of abduction in this case is that the patient's observed chills and the pain in his right upper quadrant are explained by an inflammatory abscess in his liver.

The "merge" step in the abduction procedure identifies the two boxed hypotheses shown in Table 3. In other words, the assumption is made that these two queries or hypotheses are identical, and they are merged using unification. Prior to merging, in the first hypothesis, the object present is known (an inflammatory abscess) but its location is unknown. In the second hypothesis, the location is known (the liver) while the object is unknown. When the hypotheses are unified, they share information. After merging, all variables are bound and the two hypotheses become one. In the end, the chills and pain are explained as a

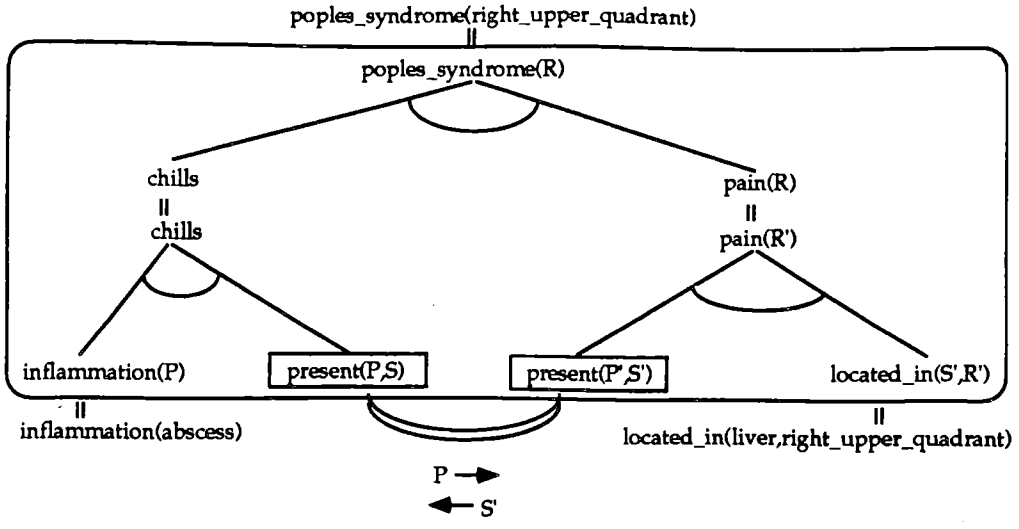

Figure 2: An Explanation for Pople's Liver Diagnosis Example

result of the presence of an inflammatory abscess in the liver.

Figure 2 shows the structure of the resulting explanation. The root nodes in the forest of proof trees are the observed symptoms (in this case, pain in the right upper quadrant and chills). The leaves are the hypotheses and facts that explain the symptoms by way of general rules (in this case, an inflammatory abscess is present in the liver). Unifications are represented as equalities in the figure. The information sharing due to merging is indicated in the figure using arrows: the value of P propagates to the right while the value of S' propagates to the left.

In the final step in the construction of the explanation, the abduction engine assumes that an abscess is present in the liver in order to complete the explanation. This was not given as part of the general background knowledge, nor was it provided as part of the statement of the case. It is also not operational, in the sense that it is not easy to directly observe whether an abscess is present in someone's liver.[6]

Using an abduction engine for the explanation component of an EBL system yields several advantages with respect to learning. First, explanations are possible in situations where proofs are not. This enables the system to learn in additional situations, using traditional explanation-based macro-learning. Second, abduction is a form of knowledge level learning. The hypotheses that are assumed as part of the process of completing explanations provide new information that is useful in its own right.

In the liver example, the abduction engine generates and learns a hypothesis stating that an abscess is present in the liver. This is a relatively specific assertion that bears on

[6]See Section 4.2 for further discussion of this point.

Table 4: A Macro Learned from Pople's Liver Diagnosis Example

$$poples\_syndrome(R) \leftarrow present(I, S_1) \wedge inflammatory(I) \wedge present(P, S_2) \wedge located\_in(S_2, R).$$

Table 5: A Special Case of the Learned Macro

poples_syndrome(R) ← *inflammatory(P),present(P,S),located_in(S,R)*

the case at hand but it is still important because it can be used to generate observable predictions and to suggest treatments.

The macro learned in this example is shown in Table 4. This macro corresponds to the enclosed part of the explanation structure in Figure 2. The macro results from ignoring the details of the particular example.

Note that the two *present* predicates that were merged in the example are not merged in the learned macro. Merging at learning time would produce the relatively compact macro shown in Table 5. However, merging at learning time is more dangerous than merging at explanation time in the sense that it is more likely to lead to errors and contradictions (see Section 3.3). This is because the literals in the condition of the macro are so general and there are so many unbound variables. Many merge errors can be avoided by considering merging only literals that were successfully merged in the example. But merging at learning time is ultimately unnecessary since merging will occur when the macro is used to construct new explanations. The unmerged macro is more general and contains the merged macro as a special case. An advantage of learning the unmerged macro is that if merging specializes it for a new example and a merge fails, it is still possible to retract the merge and use a more general form of the macro.

3 Case Studies

We studied the integration of abduction and learning on numerous examples in the context of diverse domains. We chose domains that enabled us to study explanations involving logical, physical, and psychological knowledge and reasoning. We chose domains relevant to a wide range of tasks including:

- natural language understanding and plan recognition,

- qualitative physical reasoning and postdiction,
- diagnosis and signal interpretation, and
- decision-making under uncertainty.

In this section, we provide brief descriptions of the case studies. Each description shows how to set up an abduction engine in a particular domain and how abduction can solve a particular task in that domain. We show how to provide and how to represent the relevant knowledge. We specify the operationality and assumability criteria and the constraints imposed on the search for explanations. We give an example and its explanation. We describe what is learned by macro-learning and by abductive inference in each case study.

3.1 Case Study: Explaining Emotions

The domain in this case study is emotions. The task is postdiction. In this case, we studied explanations of emotional states in terms of prior situations and events. This task is relevant to natural language understanding and plan recognition (Dyer, 1983a; Dyer, 1983b).

We constructed a first order logical theory of emotion elicitation containing rules covering eliciting conditions of twenty emotion types (see Table 6). In addition, we coded variants of a number of them, details of which have been omitted due to space constraints. (See O'Rorke & Ortony, 1992 for a presentation of the full theory.) The theory draws upon knowledge representation work on situation calculus (McCarthy, 1968) and conceptual dependency (Schank, 1972). It includes axioms that support causal reasoning about actions and other events that can lead to emotional reactions. For example, the first law below mediates positive and negative effects of actions. The second law states that a precondition of a physical transfer from one location to another is that one must first be at the initial location. The remaining laws state the effects of a physical transfer.

$$holds(F, do(A, S)) \leftarrow causes(A, F, S) \wedge poss(A, S).$$
$$poss(ptrans(P, To, From, T), S) \leftarrow holds(at(T, From), S).$$
$$causes(ptrans(P, To, From, T), at(T, To), S)$$
$$causes(ptrans(P, To, From, T), \overline{at(T, From)}, S).$$

Emotion types are represented as fluents and their eliciting conditions are encoded in rules. As examples, consider the rules for the emotion types *fear* and *relief*, shown in Table 6. The *fear* rule captures the idea that people may experience fear if they want an anticipated fluent not to hold. Relief may be experienced when the negation of a feared fluent holds. Fear usually occurs before the fluent holds. Note that, although many examples of fear involve expectations, we use the predicate *anticipates* in an effort to suggest the notion of "entertaining the prospect of" a state of affairs. The purpose of this

Table 6: Elicitation Conditions for 20 Emotion Types

$$
\begin{aligned}
joy(P,F,S) &\leftarrow wants(P,F,S) \wedge holds(F,S).\\
distress(P,F,S) &\leftarrow wants(P,\overline{F},S) \wedge holds(F,S).\\
happy\_for(P_1,P_2,F,S) &\leftarrow joy(P_1,joy(P_2,F,S_0),S).\\
sorry\_for(P_1,P_2,F,S) &\leftarrow distress(P_1,distress(P_2,F,S_0),S).\\
resents(P_1,P_2,F,S) &\leftarrow distress(P_1,joy(P_2,F,S_0),S).\\
gloats(P_1,P_2,F,S) &\leftarrow joy(P_1,distress(P_2,F,S_0),S).\\
hopes(P,F,S) &\leftarrow wants(P,F,S) \wedge anticipates(P,F,S).\\
fears(P,F,S) &\leftarrow wants(P,\overline{F},S) \wedge anticipates(P,F,S).\\
satisfied(P,F,S) &\leftarrow precedes(S_0,S) \wedge hopes(P,F,S_0) \wedge holds(F,S).\\
fears\_confirmed(P,F,S) &\leftarrow precedes(S_0,S) \wedge fears(P,F,S_0) \wedge holds(F,S).\\
relieved(P,\overline{F},S) &\leftarrow precedes(S_0,S) \wedge fears(P,F,S_0) \wedge holds(\overline{F},S).\\
disappointed(P,\overline{F},S) &\leftarrow precedes(S_0,S) \wedge hopes(P,F,S_0) \wedge holds(\overline{F},S).\\
proud(P,A,S) &\leftarrow agent(A,P) \wedge holds(did(A),S) \wedge praiseworthy(A).\\
self\_reproach(P,A,S) &\leftarrow agent(A,P) \wedge holds(did(A),S) \wedge blameworthy(A).\\
admire(P_1,P_2,A,S) &\leftarrow agent(A,P_2) \wedge holds(did(A),S) \wedge praiseworthy(A).\\
reproach(P_1,P_2,A,S) &\leftarrow agent(A,P_2) \wedge holds(did(A),S) \wedge blameworthy(A).\\
grateful(P_1,P_2,A,S_1) &\leftarrow agent(A,P_2) \wedge holds(did(A),S_1) \wedge precedes(S_0,S_1) \wedge\\
&\quad causes(A,F,S_0) \wedge praiseworthy(A) \wedge wants(P_1,F,S_1) \wedge holds(F,S_1).\\
angry\_at(P_1,P_2,A,S_1) &\leftarrow agent(A,P_2) \wedge holds(did(A),S_1) \wedge precedes(S_0,S_1) \wedge\\
&\quad causes(A,F,S_0) \wedge blameworthy(A) \wedge wants(P_1,\overline{F},S_1) \wedge holds(F,S_1).\\
gratified(P,A,S_1) &\leftarrow agent(A,P_2) \wedge holds(did(A),S_1) \wedge precedes(S_0,S_1) \wedge\\
&\quad causes(A,F,S_0) \wedge wants(P,F,S_1) \wedge holds(F,S_1) \wedge praiseworthy(A).\\
remorseful(P,A,S_1) &\leftarrow agent(A,P_2) \wedge holds(did(A),S_1) \wedge precedes(S_0,S_1) \wedge\\
&\quad causes(A,F,S_0) \wedge wants(P,\overline{F},S_1) \wedge holds(F,S_1) \wedge blameworthy(A).
\end{aligned}
$$

is to avoid suggesting that hoped-for and feared events necessarily have a high subjective probability.

The operationality criterion in this domain enables efficient recognition of predicates such as the following.

diff(X, Y).
member(X, Y).
opposite(X, Y).
action(X).
precondition(A, C).
agent(A, P).
precedes(S1, S2).
cognitive_unit(P, Q).
d_likes(P, Q).

Some of these predicates are relatively domain independent and they will be seen in other case studies. Others are relatively domain-dependent. The predicate *diff*, which ensures that two terms are not identically equal, is needed in representing knowledge about many domains. The predicate *action* is useful in domains that involve agents and their actions. This predicate checks whether its argument is an action. It is operational since no explanatory reasoning is involved. The predicates *cognitive_unit* and *d_likes* are relatively domain dependent. The dispositional attitude *d_likes* is operational because it is assumed that likes and dislikes are not explainable. In addition to these predicates, simple goals present in the database as facts are also considered to be operational.

All conjectures in this domain are assumable except meta-predicates like *diff* and instances of the following:

$$preconditions(A, F).$$
$$causes(A, F, S).$$

In other words, the abduction engine is not allowed to assume that an arbitrary fluent might be a precondition for an action, nor is it allowed to assume unprovable cause-effect relationships between actions and fluents. In early experiments without these constraints we found that the abduction engine conjectured large numbers of implausible causal relationships. We impose these constraints because, in this domain, many explanations of given observations are possible. It is important to constrain the search to avoid large numbers of implausible hypotheses and explanations.

We use an example to illustrate the abductive construction of explanations for emotions. The example is based on data taken from a diary study of emotions (Turner, 1985). Most of the subjects who participated in the study were sophomores at the University of Illinois at Champaign-Urbana. They were asked to describe emotional experiences that occurred

Table 7: Explanations for Relief and Fear

Case Facts
wants(mary, sleep(mary), _)
Query
why(relieved(mary, not at(tc, home(mary)), s2))
Explanation
relieved(mary, not at(tc, home(mary)), s2)
precedes(s1, s2)
fears(mary, at(tc, home(mary)), s1)
wants(mary, not at(tc, home(mary)), s1)
anticipates(mary, at(tc, home(mary)), s1)
holds(not at(tc, home(mary)), s2)
causes(ptrans(tc, _29887, home(mary), tc), not at(tc, home(mary)), s1)
poss(ptrans(tc, _29887, home(mary), tc), s1)
holds(at(tc, home(mary)), s1)
not causes(ptrans(karen, home(mary)), not at(tc, home(mary)), s0)
holds(at(tc, home(mary)), s0)
poss(ptrans(karen, home(mary)), s0)
Abbreviations
s1=do(ptrans(karen, home(mary)), s0)
s2=do(ptrans(tc, _591, home(mary), tc), s1)

within the previous 24 hours. They typed answers to a computerized questionnaire containing questions about which emotion they felt, the event giving rise to the emotion, the people involved, the goals affected, and so on. Over 1000 descriptions of emotion episodes were collected, compiled, and recorded on magnetic media. We have encoded over 100 of these examples using our situation calculus representation language. The following case provides examples of *relief* and *fear*.

> Mary wanted to go to sleep.
> Karen returned.
> T.C. finally left her place.
> Mary was relieved.

The case is encoded as shown in Table 7. The case fact says that Mary wants sleep. The query asks why Mary is relieved that T.C. is not at her home in the situation that results after T.C.'s departure. T.C.'s departure occurred in the situation resulting from Karen's return. (Note the abbreviations for relevant situations at the bottom of the Table.)

The explanation shown in Table 7 was constructed automatically by AMAL by backward chaining on observations to be explained. AMAL tries to reduce the observation to known facts by invoking general laws (e.g., causal laws of situation calculus and laws of emotion elicitation). In this case, the eliciting condition for *relief* is invoked in order to explain Mary's relief. This generates new questions that must be answered, and so on. The resulting explanation (shown in Table 7) states that Mary is relieved that T.C. is no longer at her home. The explanation assumes that Mary fears T.C.'s presence in her home because she wants T.C. not to be in her home but she anticipates that he will be there.

The following macro is learned from the example.[7]

$$\begin{aligned} &relieved(P_1, \overline{P}, do(ptrans(P_2, T, F, O), S_1)) \leftarrow \\ &\quad precedes(S_0, do(ptrans(P_2, T, F, O), S_1)) \wedge \\ &\quad wants(P_1, \overline{P}, S_0) \wedge anticipates(P_1, P, S_0) \wedge \\ &\quad causes(ptrans(P_2, T, F, O), \overline{P}, S_1) \wedge holds(at(O, P), S_1). \end{aligned}$$

A gloss of the macro is: a person may experience relief that a fluent does not hold in the situation resulting from a physical transfer if the person desired the fluent not to hold in an earlier situation but anticipated it holding and the move causes the fluent not to hold. The macro bypasses the intermediate emotion type *fears*, going directly to the eliciting conditions $wants(P_1, \overline{P}, S_0)$ and $anticipates(P_1, P, S_0)$. The condition that the negative fluent holds is also elaborated. It is replaced by a precondition of the action that causes the negative fluent to hold. This macro is quite general as compared to the given case. In

[7]Various macros with more or less generality can be learned here. The macro shown is derived by excising an instance of a negative frame axiom that established the preconditions for the *ptrans* in the example. A still more general macro can be learned, by replacing the *ptrans* with an arbitrary action.

addition to eliminating the specific agents involved, the macro does not require the desired negative fluent to be a direct effect of the *ptrans*, although in the specific example this was the case since Mary desired T.C. to be absent and his leaving had this effect.

An interesting general class of macros is learned in this domain if the eliciting conditions of the compound emotions *gratitude, anger, gratification,* and *remorse* are initially expressed in terms of their components. These emotion types are compounds of well-being and attribution emotion types. The compounds are formed by crossing the reactions to events and states (*pride* and *distress*) with the reactions to agent's actions (*admiration, reproach, pride, shame*). The eliciting conditions of the compound emotions shown in Table 6 were constructed by combining the eliciting conditions of the components, simplifying by eliminating redundant conditions. This approach was taken to avoid suggesting that the components are necessarily elicited whenever a compound emotion is elicited. But if the eliciting conditions of compounds such as *angry_at* are initially expressed in terms of their components, macro-learning "compiles out" the component emotions, so that compounds are explained and recognized directly in terms of the eliciting conditions of the components rather than indirectly through the component emotion types. It is interesting to speculate whether this sort of chunking occurs in human reasoning about emotions.

Another general class of macros corresponding to frame axioms is learned in this domain when explanations of effects of actions are demanded and when learning from subgoals is allowed. This corresponds to Kowalski's (1979) observation that early versions of frame axioms that were specific to individual actions can be had by forming macros from very general frame axioms plus specific statements about effects of actions.

Merging in this domain occurs primarily in the eliciting conditions of the compound emotions. *Angry_at*, for example, combines the eliciting conditions of *distress* and *reproach*. The *holds* conditions in the eliciting conditions for *angry_at* are identified when a person is *angry_at* another because the other is assumed to have done a *blameworthy* action that the person did not want done. When the *holds* conditions are not identical, they are effects of hypothetical causes that are identified. This is the case when a person is *angry_at* another because of an unpleasant effect of a *blameworthy* action presumed executed by the other. More merging occurs if the eliciting conditions of compounds such as *angry_at* are expressed in terms of their components since many redundant conditions were eliminated by hand in constructing the eliciting conditions of the compound emotions shown in Table 6. In the alternative approach, these hand-merged conditions would be merged automatically at explanation time.

Assumptions are required in the majority of the cases based on the diary study data just as they were needed in the example of *relief* and *fear*. The kinds of assumptions needed include missing preconditions, goals, prospects, and judgements. In the example, the assumption that T.C. was at Mary's home in the initial situation helps explain the fact that he was there after Karen came home. This in turn is a precondition for T.C.'s leaving Mary's home. The example also requires an assumption that Mary wanted T.C. to

be elsewhere in order to explain Mary's fear that T.C. would be at her home. Assumptions about others' goals occur frequently, especially in explaining emotions that involve the "fortunes of others." Abductive assumptions about other mental states include assumptions about whether agents anticipate events. In the example of *relief*, it was assumed that Mary anticipated T.C.'s continued (unwelcome) presence in her home. Assumptions about judgements of blame worthiness and praise worthiness are important in explaining a number of emotions not present in the example.

Operational predicates are assumable in this domain. For example, ordinarily dispositional liking is determined by looking up facts of the form *d_likes(John,Mary)* in the database in order to determine whether John likes Mary. But if no such fact is present, instead of failing the query, the query may be assumed if necessary in order to complete the explanation. In one example in the case study, this is done in order to explain why John is angry at someone who insulted Mary.

3.2 Case Study: Explaining Physical Processes

The domain in this case study is a qualitative process theory (Forbus, 1984) of chemical reactions. We used this domain to study dramatic changes in systems of beliefs such as occur in scientific revolutions. We investigated a general approach to theory revision using abduction for theory formation. The approach is based on the view that, when an anomaly is encountered, explanations of the anomaly can lead to new hypotheses that can form crucial parts of a revised theory. We studied this approach using examples based on events that occurred during the Chemical Revolution (Conant, 1957; Guerlac, 1961; Ihde, 1980). In earlier work, Thagard (1988; 1989) showed how one can choose between the phlogiston and oxygen theories, assuming that both theories are available to choose from. We extended earlier work by investigating abductive approaches to generating theories such as the oxygen theory. The specific task in this context was postdiction, the explanation of an observation in terms of previous events.

We constructed a physical domain theory based on qualitative process theory (Forbus, 1984). Our logical interpretation of QP theory included the laws shown in Table 8. The first law, GL_1, states that active processes directly influence quantities, driving them up or down. The laws GL_{2a} and GL_{2b} cover indirect influences mediated by qualitative proportionality. The first law of this pair, GL_{2a}, states that a quantity may increase or decrease if it is positively qualitatively proportional to another quantity and the other quantity increases or decreases, respectively. The second law of this pair states that a quantity may increase or decrease if it is negatively qualitatively proportional to another quantity and the other quantity decreases or increases, respectively. In other words, positive qualitative proportionalities transmit the signs of changes in proportional quantities while negative proportionalities invert the signs of changes. The law GL_3 establishes qualitative proportionalities between sums and their addends.

Table 8: Laws of Qualitative Process Theory

GL_1 : $deriv\_sign(Q_1, Sign) \leftarrow process(Process) \wedge active(Process) \wedge influence(Process, Q_1, Sign).$

GL_{2a} : $deriv\_sign(Q_1, Sign) \leftarrow qprop(Q_1, Q_2, pos) \wedge deriv\_sign(Q_2, Sign).$

GL_{2b} : $deriv\_sign(Q_1, Sign_1) \leftarrow qprop(Q_1, Q_2, neg) \wedge deriv\_sign(Q_2, Sign_2) \wedge opposite(Sign_1, Sign_2).$

GL_3 : $qprop(Q, Q_i, pos) \leftarrow qty\_eq(Q, qty\_sum(Qs)) \wedge member(Q_i, Qs).$

Table 9: A Qualitative Theory of Combustion and Calcination

$qprop(weight(P), amount(P), pos).$

$process(combustion).$

$influence(combustion, amount\_of\_in(phlogiston, charcoal), neg).$

$process(calcination).$

$influence(calcination, amount\_of\_in(phlogiston, mercurius\_calcinatus), neg).$

$qty\_eq(amount(C), qty\_sum(Qs)) \leftarrow complex(C) \wedge amounts\_of\_components\_of(Qs, [C_1, C_2|Cs], C).$

$amounts\_of\_components\_of([amount\_of\_in(C_1, C) \wedge amount\_of\_in(C_2, C)|As], [C_1, C_2|Cs], C) \leftarrow component(C_1, C) \wedge component(C_2, C) \wedge diff(C_1, C_2) \wedge components(Cs, C) \wedge amounts\_of\_components\_of(As, Cs, C).$

$amounts\_of\_components\_of(As, Cs, C) : -components(Cs, C) \wedge setof(amount\_of\_in(C_i, C), component(C_i, C), As).$

$components(Cs, C) : -setof(C_i, component(C_i, C), Cs).$

We provided a theory of combustion and calcination reflecting the phlogiston theory (see Table 9). The theory is intended to capture the following qualitative chemical ideas:

- The weight of an object is qualitatively proportional to the amount.
- Combustion is a negative influence on the amount of phlogiston in charcoal.
- Calcination is a negative influence on the phlogiston in mercurius calcinatus.
- The amount of a complex substance equals the sum of the amounts of the components.

The operationality criterion in this domain enables efficient recognition of predicates such as the following.

diff(X, Y).
member(X, Y).
setof(X, Y, Z).
opposite(X, Y).
components(Cs, C).
amounts_of_components_of(As, Cs, C).

In addition, simple goals present in the database as facts are considered to be operational.

In this domain, the system is instructed to reject assumptions of the following form:

active(P).

In other words, the system is not allowed to assume that processes are active. All other hypotheses are subject to the approval or disapproval of the user. In general, additional search control is required in this domain; depth first search tends to get stuck generating useless branches that happen to occur early in the search. In a related study (O'Rorke, Morris, & Schulenburg, 1990), we used a heuristic measure of the quality of partial explanations and conducted a best-first search. The measure was similar to that used in weighted abduction (Stickel, 1988); it favored explanations that grounded out more queries in case facts. But we found it useful to introduce an additional penalty to discourage the introduction of unnecessary individuals (Skolem constants).

Given the anomalous observation that the weight of mercurius calcinatus increases during calcination, AMAL constructs the explanation shown in Table 10.[8] The explanation is interpreted: the weight of mercurius calcinatus increases because it is qualitatively proportional to the amount. This in turn increases because it is proportional to the amount of its components since it is a complex substance and the amount of a component is increasing. The component is increasing in amount because a process, namely calcination, is actively driving the amount of the component up.

[8]The explanation shown is generated by AMAL using a depth bound that prevents runaway depth first search.

Table 10: Explanation of an Increase in the Weight of Mercurius Calcinatus

Case Facts
active(calcination)
Query
why(deriv_sign(weight(mc), pos))
Explanation
deriv_sign(weight(mc), pos)
qprop(weight(mc), amount(mc), pos)
deriv_sign(amount(mc), pos)
qprop(amount(mc), amount_of_in(_49, mc), pos)
qty_eq(amount(mc), qty_sum(amounts))
complex(mc)
amounts_of_components_of(amounts, [_49, _51| _53], mc)
component(_49, mc)
component(_51, mc)
diff(_49, _51)
components(_53, mc)
amounts_of_components_of(_52, _53, mc)
member(amount_of_in(_49, mc), amounts)
deriv_sign(amount_of_in(_49, mc), pos)
process(calcination)
active(calcination)
influence(calcination, amount_of_in(_49, mc), pos)
Abbreviations
mc=mercurius_calcinatus
amounts=[amount_of_in(_49, mc), amount_of_in(_51, mc)|_52]

Table 11: Macro Learned in the Case of Mercurius Calcinatus

$$
\begin{aligned}
deriv\_sign(X, S) \leftarrow\ & qprop(X, amount(C), pos) \wedge \\
& complex(C) \wedge component(C_1, C) \wedge \\
& component(C_2, C) \wedge diff(C_1, C_2) \wedge components(Cs, C) \wedge \\
& amounts\_of\_components\_of(As, Cs, C) \\
& member(A, [amount\_of\_in(C_1, C), amount\_of\_in(C_2, C) \mid As]) \\
& process(P) \wedge active(P) \wedge influence(P, A, S).
\end{aligned}
$$

AMAL proposes several explanations prior to the one shown in Table 10. The first explanation proposed is that calcination directly influences the weight of mercurius calcinatus, driving it up directly. The second explanation proposed is that calcination influences the amount of mercurius calcinatus, driving it up and thus indirectly driving the weight up. These explanations are more or less compatible with the explanation shown in Table 10. We reject them not because they are false, but because we desire a more specific explanation taking into account the fact that a calx is a complex substance. For a discussion of this sort of preference for more specific explanations, see (Poole, 1985).

Macro learning in the example produces the results shown in Table 11. The macro is interpreted to mean that a quantity may change if it is qualitatively proportional to the amount of a complex substance and a process is actively influencing the amount of a component of that substance. This macro is more general than the example; it does not specify whether the change is an increase or decrease and it does not specify names of the complex substance or the specific components.

Merging occurs infrequently in the examples studied in this domain. When it does occur, it is sometimes spurious. For example, merging can collapse components associated with complex substances. The requirement that the components be distinct (expressed using the *diff* predicate) then causes backtracking that undoes such merging.

Assumptions are necessary in the explanation of the observation shown in Table 10. The explanation rests in part on facts of the case at hand, such as the fact that calcination occurred, and in part on general facts, such as the proportionality between weights and amounts of physical substances. But the given facts are not sufficient to construct an explanation of the observation. The explanation requires several hypotheses (shown in boxes). The hypotheses are that mercurius calcinatus is a complex substance and calcination influences the amount of a component of this substance, causing it to increase. The unknown component is a hypothetical object (a Skolem constant) invented as a natural consequence of the explanation process. This unknown substance corresponds to the theoretical entity "oxygen" invented by Lavoisier. The hypotheses associated with it correspond to crucial

parts of the oxygen theory of combustion.

Lavoisier's hypothesis that something was added by calcination to calx of mercury, in conjunction with experimental results of Priestley and others, eventually led him to posit the existence of a hitherto unknown component of air. During a period of over a decade, Lavoisier and his colleagues worked out a new theory of combustion, calcination, and respiration that eventually displaced the phlogiston theory. This occurred because most chemists of the time were persuaded that the new theory explained the new observations (and re-explained old observations) in a more coherent manner than did modified versions of the phlogiston theory.[9]

3.3 Case Study: Explaining Decisions

The task in this case study is decision-making in situations involving uncertainty and outcomes of differing utility. Such decisions typically involve trade-offs between conflicting goals. This task is both ubiquitous and general; decision-making plays an important role in many tasks such as plan recognition and it specializes to tasks such as diagnostic decision-making.

In this abstract domain, we provide background knowledge in the form of a *qualitative logic of decision* (O'Rorke & El Fattah, 1991). This is a first order logical theory with three main components: a decision theory, a theory of arithmetic inequalities, and an operationality criterion.

The decision theory specifies when actions are reasonable. Given that *op* is an action under consideration, assume that P is a predicate such that the odds of P being true is *odds*. Assume that δ_1 is the difference between the utility of the outcome of doing *op* when P is true minus the utility of the alternative to doing *op* when P is true and that δ_2 is the difference in utilities in case P is false. The following rule is an example of the sort of decision theory provided.

$$should(\delta_1, \delta_2, odds) \leftarrow \delta_2 > -odds \times \delta_1.$$

This rule simply states an instance of the Bayesian view that one should prefer to do an action of maximum expected utility.

Knowledge of arithmetic inequalities enables the system to reason about constraints on quantities. Some of the more important rules provided about arithmetic inequalities are shown in Table 12. These rules provide information about arithmetic inequalities including the fact that inequalities are transitive, multiplication by non-negative numbers does not alter the direction of an inequality, and so on.

[9] We have not attempted to model the overthrow of the phlogiston theory, or the argumentation or the social processes involved. We do not claim to have automated all of the reasoning involved in the chemical revolution.

Table 12: Some Rules of Real Arithmetic

$T1:$	$X > Z \leftarrow X > Y \wedge Y > Z$
$T2:$	$X > Z \leftarrow X \geq Y \wedge Y > Z$
$MP:$	$A \times X > A \times Y \leftarrow X > Y \wedge A > 0$
$MN:$	$A \times X \geq A \times Y \leftarrow X \geq Y \wedge A \geq 0$

The operationality criterion is not just a list of operational predicates in this domain. Operationality requires explanations to be based upon direct comparisons of relevant quantities. Only comparisons between utilities and odds and qualitative landmarks are allowed. The admissible landmarks for odds are 0 and 1. The admissible landmark for utilities is 0. The utilities and odds introduced in a problem are also considered to be (problem specific) landmarks. They are admissible whether they are numeric or symbolic (variables). The negatives of all these quantities are also allowed. The use of inequalities and the representational bias provided by this operationality criterion contribute to the qualitative character of the logic of decision.

Search is controlled in this domain in part by testing new queries such as ground inequalities in order to verify their truth before trying to explain them. For example, upon generating new queries such as $gt(20, 1)$ or $gt(0, 1)$, the system verifies the first and rejects the second goal by calling PROLOG using the built-in predicate $>$. Considerable effort is saved by avoiding attempts to explain false statements.

Assumability is determined by the user. Examples fall into two classes. One class involves numeric values for all parameters and no assumptions. The second class involves unknown values and assumptions. The following is an example where assumptions provide valuable information.

Given a concrete decision problem specified as a query about whether it makes sense to do a given operation under given conditions, AMAL attempts to construct a proof that the action should be taken in the given situation. AMAL generalizes the particular decision problem by specializing the general decision theory and theory of arithmetic. Irrelevant details of the case are discarded. The underlying abstract explanation can be used to justify the current decision and to decide similar cases in the future.

Consider the following example. The B-1B strategic bomber is maintained using an on-board Central Integrated Test System (CITS). A ground-based expert system called the CITS Expert Parameter System (CEPS) was developed by Rockwell International Corporation and Boeing Military Airplane Company. This diagnostic expert system was constructed using standard knowledge-engineering techniques. The following quote is from a report on CEPS (Davis, 1986).

Table 13: The B-1B Example

operation	bad gate	bad servo	expected utility
check gate	$-1 \times p$	$-(1+12) \times (1-p)$	$12 \times p - 13$
replace servo	$-(12+1) \times p$	$-12 \times (1-p)$	$-12 - p$

> *The resolution of this ambiguity was determined by interviewing maintenance experts. CITS flags a failure in the Stability Control Augmentation System involving a bad servo assembly. This error code indicates an 8 to 16 hour job to replace the servo. However, experienced technicians have noted that sometimes the failure is actually caused by a faulty test gate that can be easily checked by a Ground Readiness Test and replaced in less than one hour. CEPS uses this expertise when it encounters this CITS Maintenance Code by instructing the technician on the appropriate test to run, and the results that he should expect.*

Evidently, experienced technicians prefer to first test a gate indicating that a servo assembly is bad instead of replacing the servo assembly first. A simplified description of the situation is shown in Table 13.

Assuming that just one of the possible explanations of the fault indication is correct, the table shows two columns — one corresponds to a faulty test gate and the other corresponds to a bad servo assembly. The rows of the table correspond to two action sequences; one sequence starts with a test of the gate and the other simply starts by replacing the servo assembly. We assume that the probability that the test light is bad is p and that the probability that the servo assembly is bad is $1 - p$. Measuring costs (negative utilities) in hours and averaging the estimated 8-16 hours yields a -12 hour cost for replacing the servo assembly. We charge -1 hour for testing and optionally replacing the test gate.

If a bad gate gives rise to the indication of a fault (an event with probability p), then checking and replacing the gate will solve the problem at a cost of one hour. If the servo is actually bad (the probability of this is $1 - p$), then we will have to replace it, too, at an additional cost of 12 hours (for a total cost of $-1 - 12 = -13$). The weighted average of the costs if we check the gate first is thus $12 \times p - 13$.

If the servo is replaced without checking the gate first, then assuming that the gate is bad and that it continues to indicate a faulty servo, we still have to replace the gate at a total cost averaging 13 hours. If the servo was actually bad as indicated, then we only lose an average of 12 hours. The weighted average of these costs is $-12 - p$ hours.

In terms of gains, losses, and odds, $\delta_1 = 12$, $\delta_2 = -1$, and *odds* is unknown in this example. The important point is that checking the test lamp first results in a substantial savings (an average of 12 hours) when the lamp is at fault. The time lost when the servo

Table 14: Explanation of the Expert's Recommended Action

```
Query
   why(should(12, -1, Odds))
Explanation
   should(12, -1, Odds)
      gt(Odds*12, 1)
         gt(Odds*12, 12)
            gt(12, 0)
            [gt(Odds, 1)]
         gt(12, 1)
```

is actually at fault is relatively small. So the correct decision is relatively insensitive to the exact quantities involved.

AMAL generates the explanation of the expert's recommendation shown in Table 14. The explanation is constructed assuming that the expert is rational (using the rules shown in Table 12). In particular, it uses transitivity of inequality and the fact that multiplying a positive number by a number greater than 1 magnifies the positive number.

The explanation requires an assumption about the expert's assessment of the odds of the lamp being the source of the problem. To be specific, the explanation includes the hypothesis that $odds \geq 1$.

Macro learning yields the following rule:

$$should(gain, loss, odds) \leftarrow gain > 0 \wedge odds > 1 \wedge gain > -loss.$$

This rule may be glossed: the operation should be done if the potential gain is greater than the potential loss and the odds of success are better than even. This is considerably more general than the example, which involved specific costs and utilities of outcomes. The rule learned from this particular decision is useful in justifying the decision in more general terms. It may also be applied to future examples, even if they are not fully specified because the exact values of some of the relevant probabilities or utilities are unknown.

Merging is especially hazardous in this domain. A small number of predicates occur very frequently, so their arguments — unrestricted variables corresponding to real numbers — tend to be identified relatively arbitrarily. This leads to explanations and macros that are inconsistent or excessively specific. This domain motivated the absence of merging in the present macro-learning method (see Section 2.2).

Abductive hypothesis formation generates conjectures about subjective assessments of likelihood in this domain (as in the example). In addition, it generates hypotheses about expert's preferences and constraints on utilities of outcomes.

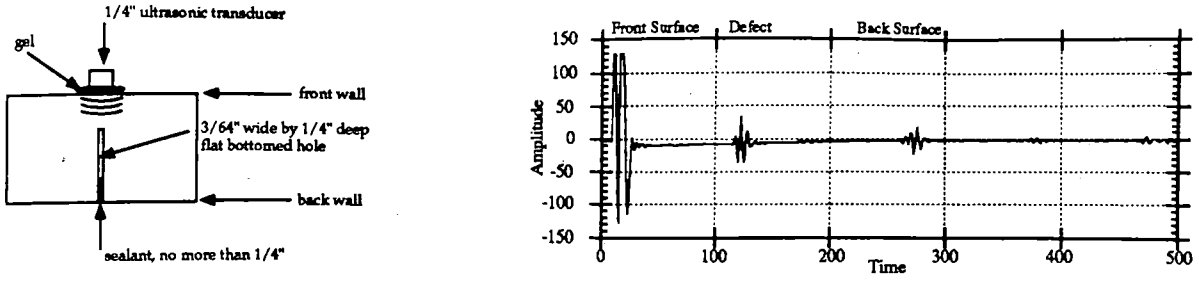

Figure 3: An NDE Test and the Resulting Signal

3.4 Case Study: Explaining Signals

The task in this case study was diagnostic signal interpretation. In particular, we studied this task in the domain of non-destructive evaluation (NDE). The goal of NDE is to determine whether parts are "good" or "defective" without damaging the parts. Piezoelectric probes are used to produce ultrasonic signals that indicate the existence of cracks and other defects inside solid materials. We use knowledge-intensive explanatory reasoning to construct interpretations of NDE signals, simultaneously classifying materials under evaluation.

We provide knowledge about the classes of interest, knowledge about the geometry of the parts to be tested, and knowledge about the physics of ultrasonic signal propagation. In a study involving a standard set of aluminum test blocks, there are two classes, *normal* and *cracked*. We provide the system with rules like the following.

$$\begin{aligned} peaks(S,3) \leftarrow\ & signal(S,B) \wedge cracked(B) \wedge \\ & feature(peak(S,1)) \wedge reflection(peak(S,1),S,front(B),B) \wedge \\ & feature(peak(S,2)) \wedge crack(C,B) \wedge reflection(peak(S,2),S,C,B)) \wedge \\ & feature(peak(S,3)) \wedge reflection(peak(S,3),S,back(Block),B). \end{aligned}$$

This rule can be glossed: the signal from a block will have three significant peaks if the block is cracked and the first peak is a reflection from the front, there is a crack in the block and the second peak is a reflection from the crack, and the third peak is a reflection from the back end.

The geometrical knowledge is restricted to one dimension: we specify the depths of known features for each of the test blocks. Knowledge of the physics of ultrasonic signals included the following rules for determining when a peak in a signal corresponds to a reflection of a feature in the material and for computing the depth of a feature from the arrival time of the reflection from the feature.

$$reflection(peak(S,N),S,F_B,B) \leftarrow signal(S,B) \wedge feature(peak(S,N)) \wedge feature(F_B) \wedge$$

$$
\begin{aligned}
& depth_c(peak(S,N), D_S) \wedge depth_l(F_B, D_B) \wedge \\
& requal(D_S, D_B).
\end{aligned}
$$

$$
depth_c(P, D) \ :- \ ptime(P, T) \wedge D = 0.00499 \times T - 0.0902
$$

Operational predicates included predicates for determining whether two numbers were sufficiently close to be "roughly equal" (*requal*), predicates for looking up the depths of known features ($depth_l$), and for computing the expected depths of features from signal reflection times ($depth_c$).

In this domain, everything is assumable subject to user approval. The user is expected to be a knowledgeable NDE technician. The system is intended to advise and not to replace the user. The abduction system is intended to be used to generate conjectures and to propose them to a human expert for ratification.

Given ultrasonic test signals, we run a preprocessor to generate qualitative abstractions of the signals specifying the times and sizes of important features such as peaks (Morris, 1992). We give AMAL queries such as "*cracked(block35)?*" and it produces explanations supporting such conclusions.

The explanation shown in Table 15 is generated for a test signal taken from a particular block. Given a qualitative description of a signal such as the one shown in figure 3, AMAL produces an explanation that identifies the first and last spikes in the signal as reflections from the front and back surfaces of the block. The explanation has subexplanations interpreting each major feature of the signal. The assumptions in boxes assert the existence of a hypothetical crack in the block at a depth roughly equal to a half inch.

Macro learning produces the rule shown in Table 16 on the given example. The rule explains three peaks in a signal taken from a block in terms of correspondences between features of the signal and the block. The rule drops intermediate concepts such as *reflection*. In addition, the macro construction process eliminates a number of duplicate *signal* and *feature* conditions that appear in the conditions of the component rules.

Assumptions about the existence of a crack are required in order to explain the anomalous peak in the signal given in the example. This involved the introduction of a hypothetical entity — a new feature of an existing object. In other examples, assumptions about the existence of echoes are required to explain anomalous spikes in signals. In general, abductive inference in this domain generates hypotheses about processes that might account for known features of signals and hypotheses about new features of existing objects.

3.5 Summary of Case Studies

We have conducted case studies involving over a hundred examples in four distinct domains. We presented sketches of these case studies and examples. We gave examples of two types of learning: abductive inference and macro learning. Here we provide a brief summary and review of the case studies prior to a discussion of the lessons learned.

Table 15: Explanation of an Anomalous NDE Signal

Case Facts
- signal(d35,block35).
  - feature(peak(d35,1)).
    - ptime(peak(d35,1),13.5).
    - amplitude(peak(d35,1),115.26).
  - feature(peak(d35,2)).
    - ptime(peak(d35,2),116.5).
    - amplitude(peak(d35,2),14.33).
  - feature(peak(d35,3)).
    - ptime(peak(d35,3),273).
    - amplitude(peak(d35,3),18.93).

Query
- why(peaks(d35, 3, block35))

Explanation
- peaks(d35, 3, block35)
  - signal(d35, block35)
  - cracked(block35)
  - feature(peak(d35, 1))
  - reflection(peak(d35, 1), d35, front(block35), block35)
    - signal(d35, block35)
    - feature(peak(d35, 1))
    - feature(front(block35))
    - depth_c(peak(d35, 1), -0.022835)
    - depth_l(front(block35), -0.02533)
    - requal(-0.022835, -0.02533)
  - feature(peak(d35, 2))
  - crack(crack0(block35), block35)
  - reflection(peak(d35, 2), d35, crack0(block35), block35)
    - signal(d35, block35)
    - feature(peak(d35, 2))
    - feature(crack0(block35))
    - depth_c(peak(d35, 2), 0.491135)
    - depth_l(crack0(block35), _1831)
    - requal(0.491135, _1831)
  - feature(peak(d35, 3))
  - reflection(peak(d35, 3), d35, back(block35), block35)
    - signal(d35, block35)
    - feature(peak(d35, 3))
    - feature(back(block35))
    - depth_c(peak(d35, 3), 1.27207)
    - depth_l(back(block35), 1.25)
    - requal(1.27207, 1.25)

Table 16: Macro Learned from the Interpretation of the NDE Signal

$$
\begin{aligned}
peaks(S,3,B) \leftarrow\ & signal(S,B) \wedge cracked(B) \wedge \\
& feature(peak(S,1)) \wedge feature(front(B)) \wedge \\
& depth\_c(peak(S,1), D_{P1}) \wedge depth\_l(front(B), D_{FB}) \wedge requal(D_{P1}, D_{FB}) \wedge \\
& feature(peak(S,2)) \wedge crack(C,B) \wedge feature(C) \wedge \\
& depth\_c(peak(S,2), D_{P2}) \wedge depth\_l(C, D_C) \wedge requal(D_{P2}, D_C) \wedge \\
& feature(peak(S,3)) \wedge feature(back(B)) \wedge \\
& depth\_c(peak(S,3), D_{P3}) \wedge depth\_l(back(B), D_{BB}) \wedge requal(D_{P3}, D_{BB}).
\end{aligned}
$$

In a study of explanations of emotions based on psychological research, we encoded a theory of emotion eliciting conditions and generated explanations of emotional states. The explanations involved hypotheses about missing preconditions, goals, prospects, and judgements.

In a study of explanations of physical processes based on the history of science, we investigated a possible role for abduction in large scale theory revision. We constructed a knowledge base encoding qualitative process theory, knowledge about sums and lists, and fragments of the phlogiston theory's view of combustion and calcination. Our abduction system constructed explanations of anomalous observations. These explanations contained new hypotheses corresponding to crucial parts of Lavoisier's oxygen theory of combustion and calcination.

In a study of explanations of decisions, we constructed a knowledge base encoding aspects of Bayesian decision theory and some knowledge of arithmetic and we generated explanations of experts' actions in uncertain situations involving goals of differing priority. The assumptions required to complete these explanations correspond to experts' preferences and subjective assessments of probabilities.

In a study of explanations of signals, we represented knowledge about the geometry of parts under nondestructive evaluation and knowledge of ultrasonic signal propagation. Qualitative descriptions of analog NDE signals are explained in terms of this given knowledge. Hypotheses about defects in the parts explain anomalous peaks in the signals. Interestingly, the separation of observations into expectations and anomalies is done *during* the explanation process. The process is a form of "layered abduction" (Josephson, 1989) in the sense that the system builds up an interpretation of the overall signal from interpretations of features in the signal. The task requires the use of predicates with some tolerance since continuous real valued numeric measurements are involved.

4 Discussion

In this section, we discuss the abductive macro learning method presented in section 2 in the light of the case studies presented in section 3. We compare abductive macro learning with earlier explanation-based learning methods. We argue that abduction provides important new capabilities. We discuss abductive hypothesis formation, assumability, operationality, search control, and macro learning.

4.1 Abductive Hypothesis Formation

The learning method described here is closely related to earlier EBL methods (e.g., see DeJong & Mooney 1986). In particular, the method is similar to the model of EBL presented by Mitchell, Keller, and Kedar-Kabelli (1986). The implementation is comparable to Kedar-Cabelli's and McCarty's PROLOG-EBG implementation of that model (Kedar-Cabelli & McCarty, 1987). The methods are similar in that they share the same basic view of explanation. They use proofs to represent explanations and they use backward chaining in constructing explanations. But the abductive model of the process of constructing explanations goes beyond deduction. When backward chaining fails, abduction is capable of making assumptions that help complete the explanation. A PROLOG-EBG program will fail to complete an explanation and will consequently fail to learn. By contrast, our method identifies hypotheses that would account for the observations, assumes them, and learns from the resulting explanation. As a result, unlike traditional EBL systems, abductive macro learners like AMAL are capable of learning at the knowledge level.

Numerous examples encountered in the case studies, including the examples chosen for presentation in this paper, illustrate the distinction between traditional EBL and our abductive extension. Given the same background knowledge and codifications of the cases provided with the observations to be explained, a purely deductive PROLOG-style interpreter will fail to find an explanation. The abductive inferences generated in these cases contribute to explanations, and they are valuable in their own right.

In the emotion domain, abductive inference is needed to construct the explanation of Mary's relief (Table 7) and many other explanations of emotion elicitation (see O'Rorke & Ortony, 1992). In general, in the emotion domain, abductive hypothesis formation generates inferences about the mental states (beliefs, desires, expectations, etc.) of the individual experiencing an emotion and of the other agents involved.

Admittedly, the emotion elicitation knowledge base could conceivably be extended so that some assumptions could be eliminated and replaced by deductive inferences. For example, if knowledge of ethics and standards of behavior could be provided, the number of assumptions in explanations requiring judgements of blame worthiness and praise worthiness could be reduced. But it is not likely that all relevant preconditions, desires, prospects, and judgements can be provided in advance.

In the domain inspired by the Chemical Revolution, abductive inference is needed to explain observations like the fact that the weight of mercurius calcinatus increases when it burns. This fact is anomalous according to the phlogiston theory but it can be explained in terms of more basic principles corresponding to qualitative process theory — provided that some assumptions are made. The assumptions correspond to key insights arrived at by Lavoisier when he realized that phlogiston did not leave combustibles as they burned, instead something (he coined the term "oxygen") combined with substances like mercurius calcinatus thus increasing their weight.

In the Chemical Revolution domain, abduction generates hypotheses about whether processes influence quantities; and hypotheses about complex substances and their components, and associated quantities such as amounts and weights. The hypotheses about influences violate closed world assumptions normally made in qualitative process theory (see section 3.6.3, "determining changes," in Forbus 1984). It is usually assumed that all changes are caused directly or indirectly by processes, all processes are known, and all direct and indirect influences are known. Quoting Forbus, "without these closed-world assumptions, it is hard to see how a reasoning entity could use, much less debug or extend, its physical knowledge." The closed-world assumptions guarantee efficient computation when the qualitative physical situation is thoroughly understood. However, the case study of the Chemical Revolution seems to indicate that this set of closed world assumptions is not appropriate when radical belief revision is necessary. Using deduction under these closed world assumptions, it is impossible to explain and learn from anomalous cases like the mercurius calcinatus example. Abduction offers a way of relaxing the closed-world assumptions while using and extending the kind of physical knowledge expressed in qualitative process theory.

In the domain of decision-making, abductive inference is needed in order to complete explanations of decisions to prefer one action over another when numbers relevant to the decisions are not known. In the B-1B example, the costs and utilities associated with the outcomes of test and replacement actions are available. The likelihoods of failure of the parts are not available but constraints on these likelihoods are generated given that the test operation is preferred to replacement. In general, abduction generates hypotheses that include constraints on subjective assessments of likelihood and on relative desirabilities of outcomes in this domain.

In the NDE domain, abductive inference generates hypotheses in an effort to explain anomalous signals. Physical processes such as echoes are invoked to explain anomalous features of signals. Hypothetical structural defects in the material under evaluation are conjectured. Classifications of materials as defective or non-defective are also conjectured. In this domain, the hypotheses generated are extremely important with respect to the immediate problem of classifying the material under evaluation.

4.2 Assumability and Operationality

In the abductive approach to explanation-based learning presented in Section 2.1, an "assumability" criterion plays an important role in learning. The assumability criterion is at once similar and distinct from operationality.

In the model of EBL presented by Mitchell, Keller, and Kedar-Kabelli (1986) the concept of operationality plays a crucial role. In that model, (quoting from page 52)

> *the task is to determine a generalization of the training example that is a sufficient concept definition for the goal concept and that satisfies the operationality criterion. Note that the notion of operationality is crucial for explanation-based generalization: if the operationality criterion were not specified, the input concept definition could always be a correct output concept definition and there would be nothing to learn! The operationality criterion imposes a requirement that learned concept definitions must be not only correct, but also in a usable form before learning is complete.*

The traditional operationality criterion is a specification of how a definition of a concept must be represented so that instances of the concept can be efficiently recognized. In many examples, concepts are deemed to be operational when they are directly observable. In the well known cup example adapted from Winston (1983), an explanation shows how the functional characteristics of a cup result from its structural characteristics. For example, a cup can contain liquids and one can hold it and drink from it because it has a handle. In this case, the operational definition of the cup is represented in terms of observable features like the physical parts of the cup and their structural relationships.

Operationality is viewed as a specification of concepts that require no explanation in the abductive macro learning method presented in Section 2.1. In the implementation, operational queries are handled by the underlying PROLOG interpreter and no record of any resulting backward chaining is retained in the explanation. In the case studies, operationality is defined using predicate based specifications such as the following.

diff(X,Y).
member(X,Y).
action(X).
precondition(A,C).
agent(A,P).
precedes(S1,S2).
cognitive_unit(P,Q).
d_likes(P,Q).

These specifications differ with respect to their range of applicability. Some, like *diff* and *member*, are domain independent; they occur in all the case studies. Others are somewhat more specialized. The predicates *action* and *precondition* and *agent* are useful in

domains that involve agents and their actions. *Action* is operational since no explanatory reasoning is involved in determining whether something is an action. The predicates *cognitive_unit* and *d_likes* are relatively domain dependent. The dispositional attitude *d_likes* is operational because it is assumed that likes and dislikes are not explainable.

Operationality is also definable in terms that require computation such as backward chaining and not just matching at explanation time. One example of this is the computation that determines that simple goals present in the database as facts are operational. Another example occurs in the case study of explanations of decisions. In that case study, comparisons on operational terms are considered to be operational. Operational terms are defined in terms of general landmarks such as 0 and 1 and in terms of problem-specific quantities like the odds of an uncertain predicate. This case study indicates that purely predicate-based specifications of operationality do not always suffice. In some cases, it is also necessary to specify constraints on the arguments of the predicates, some of which have to be evaluated at explanation time.

Assumability is like operationality in that both appear to be domain and task specific — so in the statement of the method and in our implementation, a hook is provided to enable assumability to be defined during the specification of each new domain. In some cases, a given list of predicates is checked to decide whether to assume a hypothesis. This approach is reasonable in situations (like NDE and diagnosis in general) where we know in advance what kinds of hypotheses we are looking for. This approach can be made more flexible by allowing computation to play a role in the decision about whether to make an assumption. In some domains, a human user is employed as an oracle at explanation-time and the decision is made interactively.

Assumability is unlike operationality in that they deal with different issues. One evaluates the plausibility of hypotheses not known in advance to be true or false while the other marks queries that can be efficiently evaluated and that do not require explanation. Operational statements are usually not assumable and assumable statements may not be operational in the usual sense. In the liver example, the hypothesis that there is an inflammatory abscess in the patient's liver is not immediately verifiable by direct observation. In fact, exploratory surgery would probably be required to conclusively confirm or disconfirm this hypothesis by visual inspection. Similarly, in the NDE domain, it is not directly observable or immediately verifiable whether a crack is present inside a solid metal part. In fact, the point of NDE is to use indirect inspection methods that avoid damaging parts by directly observing for defects. In the emotion domain, hypotheses about mental states occur frequently and these are even less directly observable.

4.3 Search Control

The abductive macro learning method presented in Section 2.1 employs two main kinds of search control. The first is a constraint on goals generated during backward chaining, and

the second is an additional constraint on goals that are being considered as hypotheses or possible assumptions.

The most interesting constraint on goals encountered in the case studies occurs in the study of decision-making. In this domain, many goals are tested using efficient computations that determine their truth prior to attempting to explain them.

The most interesting constraint on conjectures occurs in the emotion and Chemical Revolution domains. The following types of conjectures are disallowed:

$$preconditions(A, F).$$
$$causes(A, F, S).$$
$$active(P).$$

These constraints can be interpreted as assumptions. They are closed world assumptions to the effect that all of the preconditions for all actions are known, all of the effects of all actions are known, and all of the processes that are active are known.

If these closed world assumptions are made, large numbers of implausible hypotheses and explanations are eliminated on the examples in the case studies. However, we think it would be a mistake to generalize from this and claim that these types of conjectures should always be disallowed. The constraints exhibit an asymmetry between actions and processes; we required all the cause-effect relationships for actions to be known in the emotion study but we allowed conjectures about cause-effect relationships for processes in the chemical revolution study. It seems likely that alternative examples or alternative representations could be constructed that would require different constraints. For an example of a scenario that could require different constraints consider the emotion domain. It may be necessary to conjecture that a state is a previously unknown effect of an action taken by some person in order to explain why a victim of the action is angry_at the agent given that the effect elicits distress in the victim and reproach towards the agent. For an example of an alternative representation that could require different constraints, consider the NDE domain. We conjecture that if the NDE domain were formulated in terms of qualitative process theory, some of the explanations of examples of signals would require hypotheses to the effect that previously unknown processes (e.g., echoes) are active. We think it would be a mistake to claim that there is a single search control or assumability criterion that applies in all domains, tasks, representations, and examples.

4.4 Macro Learning

In this section we discuss the macros acquired in each case study. We discuss the macros' content, compactness, generality, their usefulness as abstract explanations, and their potential for providing speedup.

In the emotion domain, macro learning constructs rules that elaborate general conditions eliciting emotions. The macro-rules specialize these general conditions while gen-

eralizing observed examples. Two classes of macros of special significance occur in this domain, one involving relationships between emotions and another involving the causes and effects of actions. The first class of macros is formed when macro learning compiles out the components of compound emotions such as *angry_at* (a compound of *distress* and *reproach*) so that the compounds are recognized directly in terms of the eliciting conditions of their components. The second class of macros is formed when macro learning acquires frame axioms specialized to individual actions. Judging from the macros acquired in the examples in this domain, it seems clear that they could benefit from methods for generalizing the structure of explanations. For example, in the case of *relief* a substantially more general macro is obtained by generalizing the *ptrans* that occurs in the example to an arbitrary action.

In the domain of qualitative processes, macro learning constructs rules that specialize laws of qualitative physics by tying them together with facts about complex substances, sets of components, and so on. While the laws talk about general reasons why quantities change, the learned macros talk about specific causes of change. In the example of mercurius calcinatus, macro learning generates a rule explaining a change in a quantity qualitatively proportional to the amount of a complex substance in terms of a process actively changing an amount of a component of that substance.

In the domain of NDE signal interpretation, macro learning constructs rules that recognize high level features of signals using lower level features plus features of the material under evaluation. Substantial merging occurs in macros constructed from NDE examples. Although, in general, merging by unification is not allowed, elimination of duplicate (identically equal) conditions is allowed and is beneficial. In the NDE example, this operation eliminates six redundant *signal* and *feature* conditions.

In explaining decisions, macro learning constructs rules that recognize whether an action is rational using simple comparisons on parameters such as the relative utility of an action and the odds of an unknown proposition's truth. Such rules may not provide speedup in case all of the relevant numbers are known, since a simple calculation may be used to determine the expected utility of an action in this case. The rules may, however, provide speedup in situations where the precise values of the relevant numbers are not known. In addition, they are useful as abstract explanations, even in cases where all the numbers are known. Instead of justifying a decision in terms of an inequality involving some calculations, the rules provide a qualitative explanation that explains the specific case as an instance of a more general class of cases where the same action is obviously the right thing to do. This is an example of a general phenomenon — abstract explanations of concrete examples can transfer to subsequent problems even if they are incompletely specified. Furthermore, abstract explanations are useful in their own right and need not always be justified in terms of performance extensions or speed up on subsequent problems. At explanation time, they improve the comprehensibility of a classification or decision by placing the example in a larger context.

4.5 The Value of Learning

In this section, we discuss the relative value of the two different forms of learning included in abductive macro learning. In this discussion, it is important to keep in mind that abductive macro learning subsumes the traditional model of explanation-based learning presented in Mitchell, et al (1986). If assumability is defined to be false so that nothing is assumable then the implementation AMAL will behave exactly like PROLOG-EBG (Kedar-Cabelli & McCarty, 1987). So abduction provides "added value" above and beyond the normal capabilities of EBL.

We argue that the value added by abductive learning is at least as important for explanation-based learning as the value of traditional deductive macro learning by appealing to the case studies. In several case studies, the majority of the examples require abductive inferences in order to complete explanations. Macro-learning is heavily dependent on abductive learning in these case studies, in the sense that it is not possible in the absence of abduction in these examples. In addition to facilitating macro-learning, which then improves performance on future problems, abduction also contributes more directly to the problem at hand. For example, in the NDE diagnosis domain, abductive inference makes a direct contribution to the interpretation of anomalous signals and the classification of materials by generating assumptions about the physical structure of the materials under evaluation. In addition to this direct contribution, abduction contributes indirectly to NDE macro learning. Without abduction, macros would only be formed on normal signals in this domain.

In a sense, comparing abduction with macro learning is like comparing apples and oranges. They do different things. Deductive macro learning does not change the epistemic state of the system; it is aimed at providing performance speedup on subsequent problems. Abductive learning changes the epistemic state of the system. It provides new knowledge relevant to the problem at hand. It does not seem possible to argue that one type of learning is more useful than the other in an absolute sense. Both types of learning are clearly useful.

5 Related Work, Limitations, and Future Work

Since Peirce's time, a great deal of work has been done on explanations and abduction. This work has taken place in fields such as philosophy (Harman, 1965; Thagard, 1981), and psychology (Donaldson, 1986), and within AI in research areas such as automated reasoning (Reiter & de Kleer, 1987), diagnosis and expert systems (Josephson, Chandrasekaran, Smith Jr., & Tanner, 1987; Peng & Reggia, 1990; Pople, 1973) naive physics, and natural language comprehension (Charniak & McDermott, 1986; Schank, 1986). A collection of articles focusing on recent AI research on abduction is available from the author (O'Rorke, 1990b).

Recent research on the relationship between abduction and other forms of reasoning (Console, Dupre, & Torasso, 1991; Konolige, 1992) shows that there is a close relationship between abduction and an alternative consistency and minimization-based approach. It is often possible to translate abduction into the alternative approach by rewriting a logical theory adding "closure statements," for example, statements to the effect that the known preconditions or causes are the only ones. While it may be worthwhile to add such closure information to deal with special cases such as necessary preconditions, in general the abductive approach is superior because it does not require complete knowledge of causation, and causal closures need not be computed and asserted.

For papers advocating the integration of abduction and explanation-based learning, see (O'Rorke, 1988) and (O'Rorke, 1990a). A more complete account of the case study of the chemical revolution is given in (O'Rorke, et al., 1990). More complete accounts of the case study of emotion elicitation are given in (O'Rorke & Ortony, 1992a; O'Rorke & Ortony, 1992b). A fuller account of the acquisition of qualitative decision rules, is given in (O'Rorke, Elliott, El Fattah, & Shu, 1992). That report examines collections of rules acquired from sets of examples. It contains a comparison of rules acquired by AMAL vs. similar rules acquired by the inductive learning program CN2. A more complete account of the abductive approach to NDE signal interpretation is given in (O'Rorke & Morris, 1992).

Related work on abduction and learning by other authors includes the following. Falkenhainer (1990) studies qualitative physical analogies and includes analogy in a process of abductive hypothesis formation. Cohen (1992) provides a connectionist approach to learning which rules to avoid applying in order to avoid erroneous explanations.

The major limitation of the mechanization of abduction discussed here is with regard to issues that arise when there are many competing explanations. How does one avoid a combinatorial explosion of possibilities while searching for plausible explanations? How does one weigh the evidence and decide that one explanation is more plausible than another? In the abduction method described here, an "assumability" predicate is used to determine which conjectures are acceptable. This approach is "all or nothing" in the sense that statements are either assumable or not. An alternative approach involves scoring functions that assign numeric "costs" to potential (partial) explanations. Stickel's heuristic method for evaluating explanations (Stickel, 1988), while originally developed in the context of an abductive approach to natural language processing (Hobbs, Stickel, Appelt, & Martin, in press; Stickel, 1989), can be used in other domains. We adopted this "weighted abduction" method in one of our case studies. More recent probabilistic approaches (Charniak & Shimony, 1990; Poole, 1991) promise to give us an even better handle on these issues. We plan to pursue probabilistic abduction in future work.

6 Conclusion

Abductive inference is important for learning. The primary advantage of abductive inference as opposed to deduction with respect to explanation-based learning is that it allows the use of assumptions in order to complete explanations. These assumptions provide useful new information, making it possible for EBL systems to learn at the knowledge level.

We presented an explanation-based learning method that integrates abduction and deductive macro-learning. The abduction component provides knowledge-level learning while the deductive learning component provides performance speed up. The combination is synergistic; abduction makes it possible to explain and learn macros from examples that cannot be explained by a deductive explanation mechanism. The method is implemented in a system called AMAL, which has been tested in case studies involving substantially different domains.

We showed how knowledge required by the abduction engine, including various forms of causal knowledge, is represented in diverse domains. Given a new domain and an abductive task requiring working backwards from observed effects to underlying causes or reasons, the general method can be applied by

- determining an operationality criterion that marks statements that need not be explained and by
- determining an assumability criterion that specifies which conjectures are allowable as assumptions.

The case studies demonstrate that the abductive approach to explanation-based learning provides a form of learning at the knowledge level that is useful in a wide variety of domains.

Acknowledgments

Yousri El Fattah, Margaret Elliott, Steven Morris, and David Aha made significant computational and research contributions to the work described here. The study of emotions was a collaborative project with Andrew Ortony. Terry Turner provided diary study data. The NDE study was a collaborative effort involving Michael Amirfathi, Bill Bond, and Dan St. Clair. M. R. Collingwood and D. J. Hagemaier provided NDE domain knowledge and test data. Deepak Kulkarni and Kiriakos Kutulakos provided signal processing programs. Thanks to anonymous reviewers of previous drafts for comments that helped improve the quality of the presentation. In addition, interactions with Pat Langley, David Schulenburg, Tim Cain, and Stephanie Sage contributed to this work.

References

Charniak, E., & McDermott, D. V. (1986). *Introduction to Artificial Intelligence.* Reading, MA: Addison-Wesley.

Charniak, E., & Shimony, S. E. (1990). Probabilistic semantics for cost based abduction. *The Eighth National Conference on Artificial Intelligence* (pp. 106–111). Boston, MA: AAAI Press/ The MIT Press.

Cohen, W. W. (1992). Abductive explanation-based learning: A solution to the multiple explanation problem. *Machine Learning,8*(2), 167–219.

Conant, J. B. (1957). The overthrow of the phlogiston theory: The chemical revolution of 1775-1789. In J. B. Conant, L. K. Nash, D. Roller, & D. H. D. Roller (Eds.), *Harvard case histories in experimental science* (pp. 65-116). Cambridge, MA: Harvard University Press.

Console, L., Dupre, D. T., & Torasso, P. (1991). On the relationship between abduction and deduction. *Journal of Logic and Computation,1*(5), 661–690.

Davis, K. D. (1986). CEPS — B-1B diagnostic expert system. *National Aerospace and Electronics Conference.* Dayton, OH.

DeJong, G. F., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning,1*(2), 145-176.

Dietterich, T. G. (1986). Learning at the knowledge level. *Machine Learning,1*(3), 287–316.

Donaldson, M. L. (1986). *Children's explanations: A psycholinguistic study.* Cambridge: Cambridge University Press.

Dyer, M. G. (1983a). *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension.* Cambridge, MA: MIT Press.

Dyer, M. G. (1983b). The role of affect in narratives. *Cognitive Science,7*, 211–242.

Falkenhainer, B. (1990). A unified approach to explanation and theory formation. In J. Shrager, & P. Langley (Eds.), *Computational Models of Scientific Discovery and Theory Formation* (pp. 157–196). San Mateo, CA: Morgan Kaufmann.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence,24*, 85–168.

Guerlac, H. (1961). *Lavoisier — the crucial year — the background and origin of his first experiments on combustion in 1772.* Ithaca, New York: Cornell University Press.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review,74*, 88–95.

Hirsh, H. (1987). Explanation-based generalization in a logic-programming environment. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 221–227). Milan, Italy: Morgan Kaufmann.

Hobbs, J. R., Stickel, M., Appelt, D., & Martin, P. (in press). Interpretation as abduction. *Artificial Intelligence.*

Ihde, A. J. (1980). Priestley and Lavoisier. *Joseph Priestly Symposium, Wilkes-Barre, Pa., 1974* (pp. 62–91). London, England: Associated University Presses.

Josephson, J. J. (1989). A layered abduction model of perception: Integrating bottom-up and top down processing in a multi-sense agent. *Proceedings of the NASA Conference on Space Telerobotics.* Pasadena, CA.

Josephson, J. R., Chandrasekaran, B., Smith Jr., J. W., & Tanner, M. C. (1987). A mechanism for forming composite explanatory hypotheses. *IEEE Transactions on Systems, Man and Cybernetics, Special Issue on Causal and Strategic Aspects of Diagnostic Reasoning,17*(3), 445–454.

Kedar-Cabelli, S. T., & McCarty, L. T. (1987). Explanation-based generalization as resolution theorem proving. In P. Langley (Ed.), *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 383–389). Irvine, CA: Morgan Kaufmann.

Konolige, K. (1992). Abduction vs. closure in causal theories. *Artificial Intelligence.*

Levesque, H. J. (1989). A knowledge-level account of abduction. *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 1061–1067). Detroit, MI: AAAI Press/MIT Press.

Michalski, R. S., & Chilausky, R. L. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems,4*(2), 125–161.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning,1*(1), 47–80.

Morris, S. W. (1992). *Methods for feature extraction from time series data* (Technical Report TR92-17). University of California, Irvine: Department of Information and Computer Science.

O'Rorke, P. (1988). Automated abduction and machine learning. In G. DeJong (Ed.), *Working Notes of the AAAI 1988 Spring Symposium on Explanation-Based Learning* (pp. 170–174). Stanford University, Palo Alto, CA: AAAI.

O'Rorke, P. (1990a). Integrating abduction and learning. *Working Notes of the AAAI Spring 1990 Symposium on Automated Abduction.* Stanford, CA: AAAI.

O'Rorke, P. (1990b). *Working notes of the AAAI 1990 Spring Symposium on Automated Abduction* (Technical Report 90–32). University of California, Department of Information and Computer Science.

O'Rorke, P., & El Fattah, Y. (1991). *A qualitative logic of decision* (Technical Report 91-08). University of California, Department of Information and Computer Science.

O'Rorke, P., Elliott, M., El Fattah, Y., & Shu, J. (1992). *Learning qualitative decision rules* (Draft). Department of Information and Computer Science, University of California, Irvine.

O'Rorke, P., & Morris, S. (1992). Abductive signal interpretation for nondestructive evaluation. In G. Biswas (Ed.), *Conference on Applications of Artificial Intelligence X: Knowledge-Based Systems* (pp. 68–75). Orlando, FL: SPIE — The International Society for Optical Engineering.

O'Rorke, P., Morris, S., & Schulenburg, D. (1990). Theory formation by abduction: A case study based on the chemical revolution. In J. Shrager, & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 197–224). San Mateo, CA: Morgan Kaufmann.

O'Rorke, P., & Ortony, A. (1992a). Abductive explanation of emotions. In J. K. Kruschke (Ed.), *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.* Bloomington, IN: Lawrence Erlbaum and Associates.

O'Rorke, P., & Ortony, A. (1992b). *Explaining emotions* (Technical Report 92-22). Submitted for publication. Irvine: University of California, Department of Information and Computer Science.

Peirce, C. S. S. (1931–1958). *Collected papers of Charles Sanders Peirce (1839–1914).* Cambridge, MA: Harvard University Press.

Peng, Y., & Reggia, J. A. (1990). *Abductive inference models for diagnostic problem solving.* New York: Springer-Verlag.

Poole, D. (1991). Representing diagnostic knowledge for probabilistic horn abduction. *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (pp. 1129–1135). Sydney, Australia: Morgan Kaufmann.

Poole, D. L. (1985). On the comparison of theories: Preferring the most specific explanation. *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 144–147). Los Angeles, CA: Morgan Kaufmann.

Poole, D. L., Goebel, R., & Aleliunas, R. (1987). Theorist: A logical reasoning system for defaults and diagnosis. In N. Cercone, & G. McCalla (Eds.), *The Knowledge Frontier: Essays in the Representation of Knowledge.* New York: Springer-Verlag.

Pople, H. E. (1973). On the mechanization of abductive logic. *Proceedings of the Third International Joint Conference on Artificial Intelligence* (pp. 147–152). Stanford, CA: Morgan Kaufmann.

Prieditis, A. E., & Mostow, J. (1987). PROLEARN: Towards a PROLOG interpreter that learns. *Proceedings of the National Conference on Artificial Intelligence* (pp. 494–498). Seattle, WA: Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning,1*(1), 81-106.

Reiter, R., & de Kleer, J. (1987). Foundations of assumption-based truth maintenance systems: Preliminary report. *The National Conference on Artificial Intelligence* (pp. 183–188). Austin, TX: Morgan Kaufmann.

Schank, R. C. (1986). *Explanation patterns: Understanding mechanically and creatively.* Hillsdale, NJ: Lawrence Erlbaum and Associates.

Stickel, M. E. (1988). A Prolog-like inference system for computing minimum-cost abductive explanations in natural-language interpretation. *International Computer Science Conference* (pp. 343–350). Hong Kong.

Stickel, M. E. (1989). Rationale and methods for abductive reasoning in natural language interpretation. In R. Studer (Ed.), *Proceedings of the International Scientific Symposium on Natural Language and Logic* (pp. 233–252). Hamburg, Germany: Springer-Verlag.

Thagard, P. (1981). Peirce on hypothesis and abduction. In K. Ketner (Ed.), *Proceedings of the C. S. Peirce Bicentennial International Congress* (pp. 271–274). Lubbock, TX: Texas Tech University Press.

Thagard, P. (1988). *The conceptual structure of the chemical revolution* (Technical Report). Princeton University, Cognitive Science Laboratory.

Thagard, P. (1989). Explanatory coherence. *The Behavioral and Brain Sciences,12*(3), 435–502.

Turner, T. J. (1985). *Diary study of emotions: Qualitative data* (Unpublished raw data). University of Illinois, Department of Psychology.

Winston, P. H., Binford, T. O., Katz, B., & Lowry, M. (1983). Learning physical descriptions from functional definitions, examples, and precedents. *Proceedings of the National Conference on Artificial Intelligence* (pp. 433–439). Washington, D.C.: AAAI Press/MIT Press.

List of Figures

1 Inputs and Outputs of the Abduction Engine 3
2 An Explanation for Pople's Liver Diagnosis Example 8
3 An NDE Test and the Resulting Signal 25