Ph.D. proposal, Department of Computer Sciences, University of Texas at Austin, 1992
Also appears as Technical Report AI 92-179, Artificial Intelligence Lab, University of Texas at Austin, March, 1991

Belief Revision in the Context of Abductive Explanation

Siddarth Subramanian

May 12, 1992

Abstract

This proposal presents an approach to explanation that incorporates the paradigms of belief revision and abduction. We present an algorithm that combines these techniques and a system called BRACE that is a preliminary implementation of this algorithm. We show the applicability of the BRACE approach to a wide range of domains including scientific discovery, device diagnosis and plan recognition. Finally, we describe our proposals for a new implementation, new application domains for our system and extensions to this approach.

1 Introduction

Assimilation of new data into an existing knowledge base is a problem that has been approached from various different angles in the Artificial Intelligence community. One task that an agent may perform on a new datum is to explain it in terms of existing knowledge in order to maintain a structured and more coherent knowledge base. The process of making fresh assumptions in order to complete such an explanation is known as abduction and is a well studied area in logic and AI [Charniak and McDermott, 1985].

Sometimes the new knowledge to be incorporated may not be consistent with the current knowledge base. If this is the case, the agent must make changes in its knowledge base in order to make the new data compatible with existing knowledge. The process of selecting a part of the knowledge to retract in order to maintain consistency with the new data is known as belief revision [Gärdenfors, 1988; Rao and Foo, 1989].

Abduction and belief revision are closely related tasks. As we will show in the next section, when multiple explanations are possible, abduction can be viewed as a search in a space of explanations. Sometimes, explanations that look preferable early on may look less preferable later, or worse still, may contradict later observations. In order to ensure that the best explanations are generated, we may choose to generate all possible explanations so that the explanations that are preferable later are generated early and are available when they are needed. This approach, which is the one taken by the ATMS [deKleer, 1987] is computationally very expensive. Another approach, which is the one taken in [Ng, 1992], would be to keep a subset of the best-looking explanations to guide the search. This approach is not, however, guaranteed to find an explanation, since all the explanations maintained may become inconsistent with later observations. The alternative we explore in this research is to maintain fewer explanations and to alter them using belief revision if they become inconsistent. Thus belief revision can be seen as a technique for improving the efficiency of abduction.

This research examines this relationship between abduction and belief revision. We present an algorithm that combines these techniques and its implementation called BRACE (Belief Revision and Abduction in the Construction of Explanations). We also look at some important problems in device diagnosis, plan recognition and scientific discovery for which we believe our algorithm is suited.

The rest of this proposal document is organized as follows. Section 2 provides some background on the two techniques of belief revision and abduction. Section 3 describes the task that we address and section 4 describes our algorithm, first in the abstract and then in a more concrete format. Section 5 looks at results of using a preliminary implementation of the algorithm to selected problems in device diagnosis. Section 6 details the problems we propose to address as future extensions of our work. Section 7 looks at related work and finally, section 8 presents some of the conclusions of this proposal.

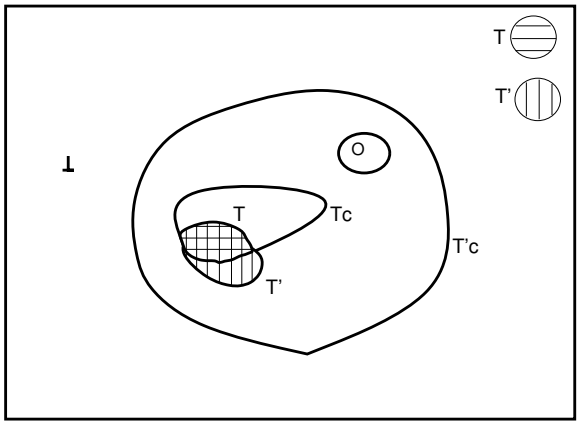

Figure 1: Abduction Task

2 Background

As mentioned in the Introduction, our research draws on two major bodies of work in the literature — work related to abduction and belief revision. The next two subsections give a general description of these two paradigms and introduce some of the related work in these areas. Further discussion of related work is left to Section 6.

2.1 Abduction

Abduction is the process of finding explanations for phenomena given background knowledge, i.e., of inferring possible causes for effects. Originally proposed by the philosopher C. S. Peirce [Peirce, 1958], abduction has been used as a reasoning mechanism for a number of applications from building explanations for natural language text [Ng and Mooney, 1990] to device diagnosis [Cox and Pietrzykowski, 1987].

The abduction task for Horn-clause theories may be defined as follows:

> **Given**: A set of Horn-clauses T (domain theory) and a set of atoms O (observations),
>
> **Find**: A minimal set of assumptions A, where $T \cup A \models O$ and $T \cup A$ is consistent.

The goal is to find a set of assumptions that, together with the background knowledge, entail or "explain" the observations. Abduction is a generalizing operator in that it involves an expansion in the belief set of the reasoner. Figure 1 is a Venn diagram representing changes to a belief set that abduction could typically cause. In this diagram, T represents the original theory while T_c represents its deductive closure. The assumptions A along with T comprise the theory T' ($T' = T \cup A$). The observations O should now be contained within the deductive closure T'_c while $\perp$ (representing inconsistency) should not.

Typically, in AI applications, this task is achieved using a backward chaining reasoner that is allowed to make assumptions where necessary [Stickel, 1988; Ng and Mooney, 1991]. We will look briefly at the ACCEL algorithm [Ng and Mooney, 1991] upon which our work is based. The algorithm is similar to the inference procedure of Prolog but with the added ability to make assumptions out of subgoals or to unify them with previous assumptions or explained atoms. ACCEL backward chains on the set of Horn-clause axioms by unifying observations with the heads of clauses. Any unexplained antecedents of clauses are queued along with other observations for backward-chaining. In addition, any antecedent (or even the observation itself) can be made an assumption.

An explanation in ACCEL is a minimal set of assumptions A that along with the theory T explains all of the observations. ACCEL does a *beam search* by maintaining some fixed number of these, preferring explanations that score higher on a specified metric such as *simplicity*. An explanation is considered simpler than some other explanation if it makes fewer assumptions. The backward chaining process continues until the queue of nodes to be explained is empty. ACCEL has been successfully applied to problems in plan recognition, disease diagnosis and device diagnosis [Ng, 1992].

2.2 Belief Revision

The problem of establishing consistency in a knowledge base faced with contradictory knowledge has been studied in various forms within philosophy, cognitive science and AI. In general, belief revision involves retracting some subset of previously held beliefs in order to accomodate the new knowledge[1]. The various theories and systems that have been studied differ in the principles used in deciding what to retract.

One dimension along which we can classify belief revision theories is in terms of what other beliefs are discarded when a particular belief is retracted. This dimension [Rao and Foo, 1989] distinguishes between systems following a Coherence Theory of Belief Revision and a Foundational Theory of Belief Revision. Systems following a coherence theory [Gärdenfors, 1988; Katsuno and Mendelzon, 1991] eliminate just enough from the original belief set to consistently accomodate the new belief whereas systems following foundational theories [Doyle, 1979; Martins and Shapiro, 1988] always discard beliefs that previously were believed only because they were consequences of other discarded beliefs.

A majority of the research involving coherence theories has been performed in the philosophy community (e.g., [Gärdenfors, 1988]). This work emphasizes minimal change in the complete deductive closure of the knowledge base as the primary criterion for determining what beliefs to retract. There is, however, no consideration of where the beliefs came from — for instance, if the statements A, $A \rightarrow B$ and B were all part of a belief set, one could conceivably retract $A \rightarrow B$ without retracting B even if the

[1] Traditionally, belief revision does include expansion and contraction operations on belief sets, but these operations are non-trivial only in the case where some beliefs need to be retracted in order to accomodate new knowledge.

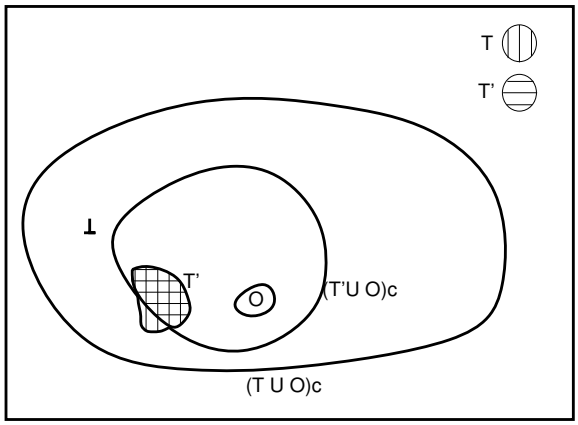

Figure 2: Belief Revision Task

only reason for believing B was that we believed A and the above implication.

Foundational theories seem to be the paradigm of choice for AI systems. In particular, the work on the Truth Maintenance System (TMS) [Doyle, 1979] and all its derivatives (for example, [deKleer, 1987]) use the idea that beliefs are only valid if they are either assumptions or they are *justified* where a justification is a set of other beliefs that together support the given belief.

All belief revision systems have some concept of minimum change: no system is biased towards making major changes in the theory when minor ones will suffice. However, they differ on what constitutes a smaller change. Another dimension of classification for belief revision systems is in terms of how they define this smallness of change. We can define minimum *semantic* change as a minimum change, using some appropriate metric, in the complete deductive closure of the knowledge base. This is generally the approach taken by philosophers and others working on coherence theories (e.g., Gärdenfors). The alternative is to use a metric for minimum *syntactic* change where we try to minimize changes to the logical sentences represented explicitly in the knowledge base. This is the approach taken by most implemented systems including the family of TMS systems.

A formal statement of the problem of belief revision would be:

> **Given**: A consistent set of beliefs comprising a theory T and a set of new beliefs O, such that $T \cup O \models \perp$,
>
> **Find**: A minimally changed theory T', $T' \subset T$, where $T' \cup O$ is consistent.

Figure 2 illustrates this problem. T represents the original set of beliefs, T_c its deductive closure, O the set of new beliefs and $(T \cup O)_c$ the deductive closure of the union of T and O. Now $\perp$ (falsity) is in this latter deductive closure and so we find a set T', $T' \subset T$, such that the deductive closure of the set $T' \cup O$ no longer includes $\perp$. Revision shrinks the set of beliefs and is thus a specialization operator.

Philosophers looking at belief revision (in particular, Gärdenfors) have come up with the idea of *epistemic entrenchment* as a guide to finding minimal revisions of the knowledge base. This is a partial ordering over beliefs that determines what is more or less likely to be retracted. Gärdenfors does not, however, commit to any particular characterization of this concept and instead makes suggestions for what it might be in different contexts. In an explanatory context, for instance, he suggests that simplicity and explanatory power ([Gärdenfors, 1988], p. 93) may be a good measure. While we do not follow his model of belief revision, our criteria for retraction are closely related to this idea.

Another distinction between systems in the literature that needs to be made to put our work in context is that between the various Truth Maintenance Systems. Our work is based on a compromise between the Justification-based TMS's (JTMS's) [Doyle, 1979] style of maintaining a single context of belief and the Assumption-based TMS's (ATMS's) [deKleer, 1987] style of maintaining all contexts. The JTMS revises contradictions using a technique called dependency-directed backtracking to resolve contradictions. The ATMS, on the other hand, maintains all possible explanations at the same time. These explanations are maintained as sets of assumptions which together support the belief. Sets of assumptions, in ATMS parlance, are known as *environments*. An environment that is inconsistent is known as a *nogood*. The ATMS deals with contradiction by keeping track of minimal nogoods and ensuring that none of them ever appear as a subset of any of the environments in the system. This eliminates the need to do revision since all the consistent sets of beliefs are already maintained by the system.

The dependency-directed backtracking algorithm in the JTMS works roughly as follows. The system looks among the justifications supporting the contradiction and constructs the set of all assumptions that support the contradiction node. It then picks culprit assumptions from among these one by one and eliminates each of them until there is no longer a valid justification for the contradiction.

Reiter [Reiter, 1987], in his work on diagnosis, proposes an algorithm for computing minimal *hitting sets* which are sets of components in a device which may be faulted to account for anomalous observations. This algorithm translates into a technique for choosing minimal sets of assumptions to retract in a JTMS type of system.

3 The Task

In this section, we motivate the task that we are addressing with an example from the domain of scientific discovery. We then describe the actual BRACE task.

3.1 An Example

The original ideas for this research came from the domain of scientific discovery and hence our motivating example is in this domain. The particular revisions we wish to model involve theories on the extinction of dinosaurs [Muller, 1988]. This problem has

also been modeled by [Thagard, 1992] although his work does not address the issues of how explanations are constructed. His methods simply evaluate explanations based on how they cohere. Our approach is to use abduction to actually construct explanations.

For a long time, it had been assumed that dinosaurs became extinct through some gradual process. However, gaps in the fossil record were discovered that made it apparent that the extinction process had been relatively sudden. This led researchers to start considering the possibility of catastrophic events causing the extinction of the dinosaurs. Among the events hypothesized were sudden climatic deterioration and huge volcanic eruptions.

Later, scientists also discovered an anomalously large deposit of iridium in the same fossil era. They also discovered that such deposits were repeated at periodic intervals of 26 million years in the fossil history and each periodic deposit was accompanied by the loss of a large number of species. Now scientists hypothesised meteor or comet strikes since meteors and comets are known to contain iridium and strikes of great magnitude could account for dinosaur extinction. Furthermore, some mechanism had to be found to explain the periodicity. One of the most popular theories for this was the idea of a twin star circling the sun that, at periodic intervals was at just the right position that comets or asteroids would get diverted from their regular paths and a collision with the earth would become more likely.

A more recent discovery made by researchers looking for an answer to the question of dinosaur extinction was the discovery by Bada and Zhao [Waters, 1990] that certain amino acids were found in anomalous amounts in the same strata of rocks as the last of the dinosaur fossils. These researchers now concluded that the comet strike hypothesis was much more likely than the meteor strike hypothesis since comets are sometimes known to have deposits of these amino acids whereas meteors are not.

What we see in this historical survey is a continual process of retraction of assumptions and re-explanation of observations. First, the gradual extinction assumption had to be dropped in favour of a sudden extinction hypothesis. Next the iridium deposits and periodicity had to be abductively explained with the twin star and meteor hypotheses. Finally, the meteor hypothesis had to be dropped in favour of a comet hypothesis.

Our algorithm thus follows the same pattern — we explain observations using abduction, retract assumptions in the face of contradictions and then attempt to re-explain observations left unexplained. This process continues until we have consistent explanations for all our observations.

3.2 Task Description

In building explanations, an agent must both expand (generalize) the set of beliefs to explain new observations and to contract (specialize) its set of beliefs to resolve contradictions. The task that this research seeks to address can be stated as follows:

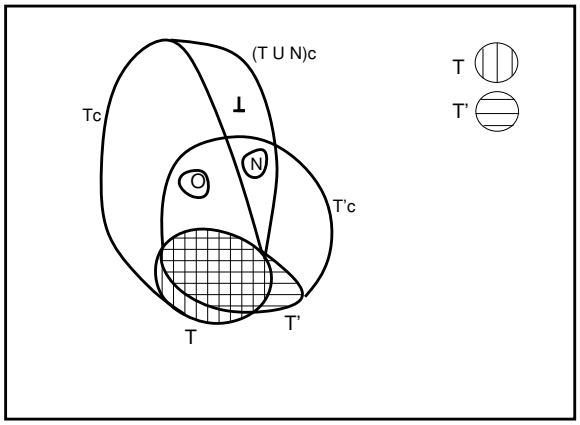

Figure 3: BRACE Task

Given: A theory T, a set of observations O such that $T \not\models O$ and $T \cup O$ is consistent, and a new set of observations N,

Find: A theory T' such that $T' \models O \cup N$ and $T \cup O \cup N$ is consistent.

Figure 3 is a Venn diagram of this task in the case where N is inconsistent with T. As before, T_c is the deductive closure of T. $\perp$ is in $(T \cup N)_c$ but not in T'_c, and this new theory T'_c includes both O and N. This combines the specialization of belief revision in removing contradictions and the generalization of abduction in encompassing observations.

If N is consistent with T, the task may just be the abduction task described earlier or revision may be involved in finding a consistent T' (i.e., $T \cup N$ may be consistent but any $T' \supseteq T$ such that $T' \models O \cup N$ may be inconsistent).

4 The Brace Algorithm

The system we propose to implement (and for which we currently have a prototype) is a first-order Horn-clause Abduction system with Belief Revision. We first look at certain broad principles underlying our algorithm and then at the details of the algorithm itself and its current implementation.

4.1 Principles

The following are some principles underlying the BRACE algorithm.

1. *Homogeneity Principle:* All knowledge in the system is encoded as assumptions.
2. *Centrality Principle:* We prefer not to retract an assumption that was previously used to explain a large number of observations.

3. *Least Cost Principle:* We prefer explanations of our observations that do not retract many of our previously held assumptions and do not make many new ones, but still explain all the previous observations.

4. *Primacy of Observations Principle:* In revising, we retract all beliefs dependent on retracted assumptions except for observations.

Typical abduction systems (e.g., [Ng and Mooney, 1991]) maintain three very distinct kinds of beliefs in their knowledge bases — *rules*, which are always believed, *assumptions* (usually in the form of atomic formulas) which are made to complete explanations, and *facts* which are always true. We claim that this distinction is not very productive. Instead, we treat all data (rules, atomic assumptions, facts etc.) as assumptions. Rules are only distinguished in that we can perform certain operations on them, such as forward and backward chaining, that we cannot perform on atomic assumptions. However, in order to perform these operations on rules, we must believe the rule itself. Our system uses a distinct proposition for each rule in the knowledge base which needs to be a part of its belief set when it is used. For instance, a rule may look like the following[2]:

$$\text{(a ?x)} \leftarrow \text{(r-1) (b ?x)}$$

The proposition (R-1) represents a belief in the rule

$$\text{(a ?x)} \leftarrow \text{(b ?x)}$$

This representation allows the system to treat belief in a rule as an atomic assumption, which is useful for implementation purposes.

By adopting this principle of *homogeneity* we have reduced our knowledge base to a form suitable for belief revision. When we perform revision we can look at all pieces of knowledge in a uniform manner as candidates for retraction.

The second principle underlying our work, *centrality*, addresses the problem of what to retract from our knowledge base when it becomes contradictory. We claim that assumptions[3] that help explain a larger number of observations (i.e., that are more central) should be preserved over assumptions that are not as useful in explanation. This corresponds closely with Gärdenfors's idea of epistemic entrenchment when the context is that of explanation.

Our algorithm uses the following definition of centrality:

$$Centrality(a) = \quad \mid O_a \mid$$

[2]Throughout this paper, we represent an atomic formula as a list whose first member is the predicate name and the other elements are the arguments. Among the arguments, a letter by itself is a constant, while one preceded by a ? is a variable.

[3]From here on, an assumption is any rule or atomic assumption in the knowledge base. Because of the homogeneity principle, these are treated uniformly.

where O_a is defined as

$$O_a = \{ \quad o : o \text{ is an observation and } a \in A \text{ where } A \text{ is a minimal, consistent set of assumptions such that } A \models o\}$$

In other words, centrality is defined to be the number of observations that a helps explain. When BRACE's knowledge base is initialized, it is given a starting set of assumptions. These are presumed to be assumptions that proved useful in explaining observations seen before the current problem solving cycle. Since these observations and other assumptions necessary to complete their proofs are not available to BRACE, initial information on centrality needs to be provided to the system. Therefore all initial assumptions must be provided with initial centralities.

The *Cost* principle is our formulation of minimal change. We prefer theories that, while explaining all of the observations, deviate least from our initial set of assumptions. A simple function to measure the deviation here is:

$$Cost = \mid A \mid + \mid R \mid$$

where A is the set of new assumptions and R is the set of retracted assumptions. The actual cost function that BRACE uses is slightly different because of the fact that some of the retracted assumptions may be initial assumptions. Since retracting initial assumptions may leave observations unexplained that are not explicitly represented in the system, we use the centrality criterion to determine Cost. The formula is thus the following:

$$Cost = \sum_{d \in R} IC(d) + \mid A \mid$$

where the function IC returns the initial centrality of its argument, and where R and A are defined as before. An assumption with a high initial centrality thus costs more to retract than one with lower centrality.

As mentioned in the previous section, belief revision systems can be classified into Foundational belief revision systems where consequents of retracted assumptions are also retracted and Coherence belief revision systems where they are not. Our work is mainly Foundational in nature with one important caveat: since theory changes must be consistent with the principal task of our algorithm which is the explanation of observed data, observations are not retracted unless directly contradicted. This is what we refer to as the *Primacy of Observations* principle. In a sense, this is consistent with foundational theory since observations have some support independent of the actual causal explanations for them. In the future work section, we look at the possibility of being able to retract observations but for the current implementation, observations are unretractable.

Another distinction mentioned in the previous section was between the ATMS style of maintaining all possible belief sets justifying a piece of knowledge and the JTMS style of maintaining just one set of beliefs. Our compromise between the two is in

```
Initialize KB with a set of initial assumptions.

Until no more observations,
        Add next set of observations to KB

        Until there is some theory that explains all observations,

                1) Forward Chain on all rules in the KB to get all
                   possible consequences.
                2) If all theories are contradictory, then REVISE.
                3) Order theories by cost and remove all but B
                   of these with least cost.
                4) ABDUCE on unexplained observations.
```

Figure 4: BRACE algorithm

using a beam search to guide our quest for explanations. We maintain a current set of theories whose size is limited by a beam width and prune all explanations that are incompatible with this set. The set is pruned by selecting theories that minimize the Cost function above.

4.2 The Algorithm and Current Implementation

In this section, we present the actual algorithm of BRACE and discuss its implementation. Since BRACE is built on top of an implementation of the ATMS, we will use some of the ATMS terminology introduced earlier.

Figure 4 is the basic algorithm of BRACE. The knowledge base is primed with a set of initial assumptions. The initially believed theory is taken to be the set of all initial assumptions. The algorithm works in batch-incremental mode i.e., it is incremental but batches of observations may be introduced at one time[4]. First, BRACE forward chains on everything in the knowledge base to assert all possible consequences. This may include the contradiction node (falsity).

At any time, a current set of alternative theories is maintained. Each theory is a set of assumptions represented in our system as deviations from the initial theory (i.e., as sets of added and retracted assumptions). A theory is contradictory if one of the environments that is a subset of it is a nogood (i.e., is a justification for (falsity)). When all current theories become contradictory, the revision routine REVISE (Figure 5) is called to propose new theories that are consistent.

[4]By some definitions of the word, this may not strictly be considered incremental since previously introduced data is maintained in the KB.

```
For every contradictory theory T,
        1) Find all minimal nogoods (contradictory sets of assumptions)
           that are subsets of T.
        2) Sort assumptions in each nogood by increasing centrality.
        3) Construct up to R sets of assumptions, in increasing order of
           total centrality, including at least one assumption from each
           nogood.
        4) Propose retractions of each of these sets and construct
           theories by retracting these sets from T.
        5) Add all observations that are no longer explained by T into
           the list of atoms to be explained.
```

Figure 5: REVISE

Once the revision phase is over, the set of theories is ordered by cost and reduced to the B theories with least cost. At this point, we enter the backward chaining abduction phase where new explanations are proposed for those observations left unexplained in some theories.

This is the algorithm at the highest level. The revision module, REVISE, is described in Figure 5. REVISE looks for contradictory sets of assumptions (nogoods) that are subsets of the theory to be revised and proposes retraction sets constructed from these including at least one assumption from each nogood. The sets are constructed in increasing order of total centrality and a limited number, R, are actually proposed. At this point, any observations that are left unexplained by the revisions are placed back in the list of atoms needing explanation.

ABDUCE, the abduction algorithm in BRACE, is described in Figure 6. First, an atom is picked from among those waiting to be expanded whose explanation is required for the largest number of the current theories. It is then expanded using all the known rules that it unifies with. All antecedents generated are introduced into the knowledge base, assumed if appropriate, and placed in the list of atoms to be explained.

The current implementation of BRACE is built on two other systems. The knowledge base is actually maintained by a modified version of COCO [Dressler and Farquhar, 1990]. COCO is itself a modification on the ATMS [deKleer, 1987] with the added ability to focus on particular environments.

ACCEL [Ng and Mooney, 1991] is an abduction system that is built on top of COCO. It uses the label propagation routines and representation of COCO but incorporates unification of first-order formulas and the beam search method described above. Explanations are compiled into a special node known as the *goal node*, which represents the conjunction of all observations. The major changes that we have made to ACCEL involve the representation of theories as changes from the initial set of assumptions,

```
Pick the node N from the list of nodes to be explained that is unexplained
in the most current theories.

For every rule R whose head unifies with N
        for each antecedent A
               1) Create a node for A' where A' is obtained from A by
                  substituting with the most general unifier of N and the
                  head of R.
               2) If A' is assumable, introduce the assumption of A' to
                  the KB.
               3) Add A' to the list of nodes to be explained.
```

Figure 6: ABDUCE

the delay in removing nogoods (since they are required for revision), the addition of a revision routine, and the computation of cost and centrality. In the next two sections, we will look at (among other things) some of the strengths and weaknesses of this implementation and changes we propose to make to the system.

5 Current Results

The preliminary implementation of BRACE described above has been tested on a simplified dinosaur extinction example as well as several examples in the domain of device diagnosis. We will first look at the former domain as an illustration of BRACE's algorithm and then at the latter as a more practical application. We will demonstrate some of the weaknesses of this implementation and motivate the proposed changes discussed in the next section.

5.1 Dinosaur Extinction: An Illustrative Example

Figure 7 shows a set of Horn-clause rules for dinosaur extinction. The predicate extinction takes two arguments, the first is the species that became extinct and the second is the approximate time period of extinction. Rules R-0 and R-1 are two different ways in which extinction phenomena could be explained. R-1 states that extinction of a species could be due to a cataclysmic-event while R-0 ascribes it to a gradual extinction of the species through that time. R-2 states that a cataclysmic event would have left a sudden gap in the corresponding fossil layer. Rules R-3 through R-5 give three types of cataclysmic events — volcanic eruptions, meteor strikes, and comet strikes.

Rule R-6 explains the phenomenon of periodic extinctions by periodic cataclysmic

Centrality			
100	(extinction ?sp ?t)	←	(r-0) (gradual-extinction ?sp ?t)
100	(extinction ?sp ?t)	←	(r-1) (cataclysmic-event ?t) (on-earth ?sp ?t)
100	(fossil-gap ?t)	←	(r-2) (cataclysmic-event ?t)
100	(cataclysmic-event ?t)	←	(r-3) (volcanic-eruption ?t)
100	(cataclysmic-event ?t)	←	(r-4) (meteor-strike ?t)
100	(cataclysmic-event ?t)	←	(r-5) (comet-strike ?t)
100	(periodic-extinctions ?t)	←	(r-6) (periodic-event ?t) (cataclysmic-event ?t)
100	(periodic-event ?t)	←	(r-7) (comet-strike ?t) (twin-star ?x ?t)
100	(periodic-event ?t)	←	(r-8) (meteor-strike ?t) (twin-star ?x ?t)
100	(deposits iridium ?t)	←	(r-9) (meteor-strike ?t)
100	(deposits iridium ?t)	←	(r-10) (comet-strike ?t)
100	(deposits amino-acids ?t)	←	(r-11) (comet-strike ?t)
100	falsity	←	(r-12) (meteor-strike ?t) (deposits amino-acids ?t)
100	falsity	←	(r-13) (gradual-extinction ?sp ?t) (cataclysmic-event ?t)
100	falsity	←	(r-14) (gradual-extinction ?sp ?t) (fossil-gap ?t)
100	falsity	←	(r-15) (volcanic-eruption ?t) (deposits iridium ?t)

Observations:
(extinction dinosaurs t1)
(fossil-gap t1)
(deposits iridium t1)
(periodic-extinctions t1)
(deposits amino-acids t1)

Figure 7: Dinosaur Extinction Theory

events.[5] R-7 and R-8 give two ways in which periodicity may arise: through either comet or meteor strikes with a twin star as a deflector.

The next three rules R-9 through R-11 describe what meteor and comet strikes deposit. Either one can explain iridium deposits while only comets can explain amino acid deposits. Rule R-12 makes it inconsistent for a meteor strike to occur with amino acid deposits.[6] Similarly, rule R-13 rules out the coincidence of a gradual extinction occurring through a period when a cataclysmic event has occurred and R-15 rules out the coincidence of iridium being deposited while a volcanic eruption has occurred. Finally R-14 states that a species that has become extinct gradually will not leave a fossil gap.

All of the rules mentioned above are introduced initially with a high centrality since they are all almost axiomatic. Therefore, when we need to revise beliefs in this domain, we are likely to look among the atomic assumptions made to explain phenomena to find assumptions to retract rather than among the rules in Figure 7.

We tested this theory on the sequence of observations shown at the bottom of Figure 7. The sequence of observations and resulting explanations are shown in Figure 8. The number of explanations maintained between sets of observations (the beam width) is a parameter to BRACE and is set, in this example, to 1. In Figure 8, theories (explanations) are represented in terms of additions and retractions of assumptions relative to the set of assumptions shown in Figure 7. The one explanation that BRACE maintains after an observation is added is shown as the singleton member of the list of current theories.

The best explanation, produced by the abduction module of BRACE for the first observation, that dinosaurs became extinct at a particular time, is that the extinction was gradual. This explanation is preferred because it involves making the fewest number of assumptions, thus minimizing the cost. When the observation of a fossil gap is introduced, this explanation is no longer consistent and BRACE retracts the assumption of a gradual extinction and replaces it with an explanation of the extinction of dinosaurs making the assumption that a volcanic eruption took place at that time.

The next observation is that of iridium deposits for the same time period. Since these are not consistent with the volcanic explanation, this assumption is retracted and an assumption of a meteor strike, causing both the extinction of the dinosaurs and iridium deposits, is made.[7]

The next observation, that extinctions are periodic phenomena, is consistent with the current explanation. However, an additional assumption, that the sun has a twin star, is made to explain the periodicity. Finally, the last observation, that deposits of

[5] The proposition (periodic-extinctions ?t) means that there are periodic extinctions with one of the extinction events occurring at time ?t. The argument to periodic-event has a similar meaning.

[6] This of course is not entirely accurate since it is actually consistent, albeit unlikely, for a meteor strike to happen at the same period of time as a comet strike. There is currently no good method of formulating a theory in a manner that avoids coincidences. This problem is discussed in the Analysis subsection of the Future Work section.

[7] This explanation actually has the same cost as that of a comet strike. The former is chosen by BRACE just because of the way the rules are ordered.

```
Observations: ((EXTINCTION DINOSAURS T1))
Current theories
(
Additions: ((gradual-extinction dinosaurs t1))
Retractions: NIL
)
Observations: ((FOSSIL-GAP T1))
Current theories
(
Additions: ((volcanic-eruption t1))
Retractions: ((gradual-extinction dinosaurs t1))
)
Observations: ((DEPOSITS IRIDIUM T1))
Current theories
(
Additions: ((meteor-strike t1))
Retractions: ((gradual-extinction dinosaurs t1) (volcanic-eruption t1))
)
Observations: ((PERIODIC-EXTINCTIONS T1))
Current theories
(
Additions: ((twin-star sk19 t1) (meteor-strike t1))
Retractions: ((gradual-extinction dinosaurs t1) (volcanic-eruption t1))
)
Observations: ((DEPOSITS AMINO-ACIDS T1)))
Current theories
(
Additions: ((twin-star sk19 t1) (comet-strike t1))
Retractions: ((gradual-extinction dinosaurs t1) (volcanic-eruption t1)
              (meteor-strike t1))
)

Run Time = 0.20 min
```

Figure 8: Sequence of Explanations for Dinosaur Observations

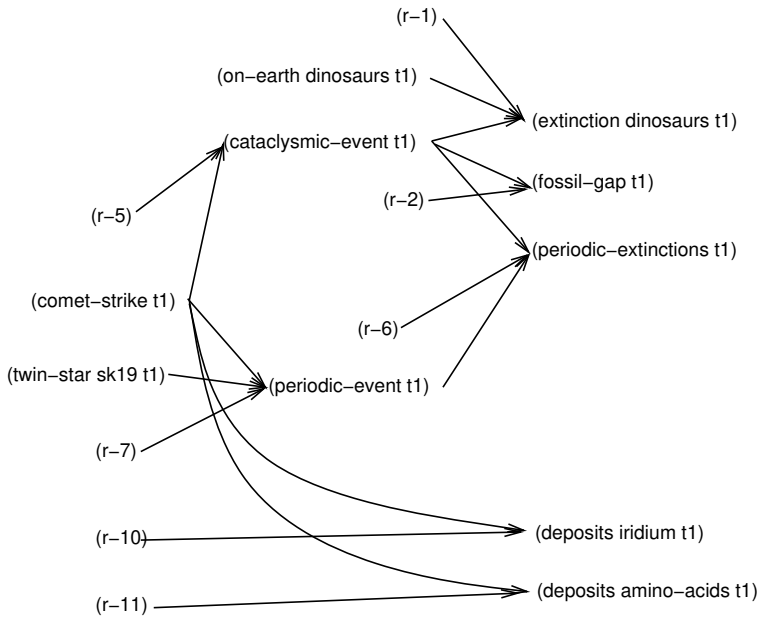

Figure 9: Explanation Graph for Dinosaur Example

certain rare amino acids are present from the same time period, causes the meteor strike explanation to become inconsistent and to be replaced by the comet strike assumption. The final explanation graph is shown in Figure 9.

5.2 Model-Based Diagnosis

The other domain on which we have tested the preliminary implementation of BRACE is the diagnosis of logic circuits. According to Poole [Poole, 1989], model-based diagnosis (where a device's model is compared to its observed behaviour and analyzed for faults) can take one of two approaches: a *consistency-based* approach or an *abductive* approach. In the former, the idea is to find minimal sets of faults in such a way that the resulting diagnostic model is consistent with the device's behaviour. This is the approach taken by most researchers in the area [Reiter, 1987; Davis, 1984; deKleer and Williams, 1987]. The alternative, an abductive approach, is taken by Ng [Ng, 1992]. Here, the idea is that the diagnosis actually explains the observed behaviour of the device. Our application of BRACE to diagnosis yields a mix of these ideas: we obtain an abductive diagnosis but only after we have located faults using a consistency-based approach.

Figure 10 shows the circuit of a full adder [Reiter, 1987]. A domain theory modeling this circuit is shown in Figure 11. The rules R-1 through R-12 model normal behaviours of different types of gates in the circuit for different combinations of inputs. Since these are definitional, they are given very high centrality. Rules R-13 and R-14 model abnormal behaviours of the circuit. We assume here that two possible kinds of abnormality are that the gate is stuck low (`ab0`) and that the gate is stuck high (`ab1`)[8].

[8]The inclusion of (`obs ?t`) as an antecedent in rules R-13 and R-14 is to allow the forward chainer

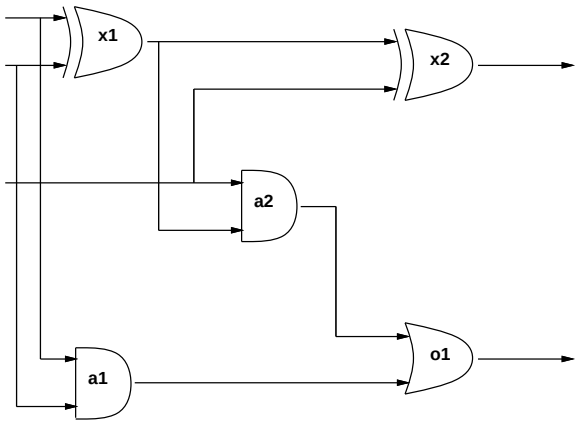

Figure 10: A Full Adder

Rules R-15 through R-17 specify the various types of gates. Rules R-18 through R-24 encode the connections between gates. These are initialized with lower centralities since it is unlikely that they have been used to explain a number of observations. On the other hand, the fact that we have made these connections does give us some confidence in the assumption that the rules hold.

The next 14 rules, R-25 through R-38, are fault rules for all gate inputs that are dependent on connection rules R-18 through R-24. Again, the fault rules indicate that the inputs may be stuck at 0 or that they may be stuck at 1. Since we initially believe the connections to be normal, these rules are not believed yet and their centralities are set at 0. The abduction system may assume these rules when they become necessary.

The rules R-39 through R-47 are the various consistency rules needed to ensure that we do not assume two different inputs at the same port, two different gate types for one gate or that a gate is both normal and abnormal at the same time. In addition to these rules, we also specify a set of initial assumptions representing our belief that the device is working correctly. The first set of assumptions says that all the gates are normal while the second set specifies the gate types.

A series of experiments were run on BRACE using this domain theory. In order to simulate faulty behaviour, the domain theory was perturbed in various ways, making assumptions of abnormal gates and broken connections, and complete sets of expected outputs for the eight complete sets of inputs were generated. The BRACE implementation was run using the rules shown in Figure 11 with outputs as observations and the inputs among the initial assumptions. The explanations of the observations constitute diagnoses of the circuits.

One additional kind of fault that was simulated was a gate-type error where x2 was implemented as an or-gate instead of an exclusive-or gate. This is a plausible type of fault which could occur when the wrong chip is accidentally used in a circuit. The centralities on all of the initial gate type assumptions were correspondingly decreased

to be able to conclude the out observation for a particular time. We ensure that the assumption (obs ?t) is established for every time instance for which there are observations.

Centrality			
100	(out ?x 0 ?t)	←	(r-1) (andg ?x) (in1 ?x 0 ?t) (in2 ?x 0 ?t) (norm ?x)
100	(out ?x 0 ?t)	←	(r-2) (andg ?x) (in1 ?x 0 ?t) (in2 ?x 1 ?t) (norm ?x)
100	(out ?x 0 ?t)	←	(r-3) (andg ?x) (in1 ?x 1 ?t) (in2 ?x 0 ?t) (norm ?x)
100	(out ?x 1 ?t)	←	(r-4) (andg ?x) (in1 ?x 1 ?t) (in2 ?x 1 ?t) (norm ?x)
100	(out ?x 0 ?t)	←	(r-5) (org ?x) (in1 ?x 0 ?t) (in2 ?x 0 ?t) (norm ?x)
100	(out ?x 1 ?t)	←	(r-6) (org ?x) (in1 ?x 0 ?t) (in2 ?x 1 ?t) (norm ?x)
100	(out ?x 1 ?t)	←	(r-7) (org ?x) (in1 ?x 1 ?t) (in2 ?x 0 ?t) (norm ?x)
100	(out ?x 1 ?t)	←	(r-8) (org ?x) (in1 ?x 1 ?t) (in2 ?x 1 ?t) (norm ?x)
100	(out ?x 0 ?t)	←	(r-9) (xorg ?x) (in1 ?x 0 ?t) (in2 ?x 0 ?t) (norm ?x)
100	(out ?x 1 ?t)	←	(r-10) (xorg ?x) (in1 ?x 0 ?t) (in2 ?x 1 ?t) (norm ?x)
100	(out ?x 1 ?t)	←	(r-11) (xorg ?x) (in1 ?x 1 ?t) (in2 ?x 0 ?t) (norm ?x)
100	(out ?x 0 ?t)	←	(r-12) (xorg ?x) (in1 ?x 1 ?t) (in2 ?x 1 ?t) (norm ?x)
100	(out ?x 0 ?t)	←	(r-13) (gate ?x) (obs ?t) (ab0 ?x)
100	(out ?x 1 ?t)	←	(r-14) (gate ?x) (obs ?t) (ab1 ?x)
100	(gate ?x)	←	(r-15) (andg ?x)
100	(gate ?x)	←	(r-16) (org ?x)
100	(gate ?x)	←	(r-17) (xorg ?x)
6	(in1 a1 ?x ?t)	←	(r-18) (in1 x1 ?x ?t)
6	(in2 a1 ?x ?t)	←	(r-19) (in2 x1 ?x ?t)
6	(in1 x2 ?x ?t)	←	(r-20) (out x1 ?x ?t)
6	(in2 x2 ?x ?t)	←	(r-21) (in1 a2 ?x ?t)
6	(in2 a2 ?x ?t)	←	(r-22) (out x1 ?x ?t)
6	(in2 o1 ?x ?t)	←	(r-23) (out a2 ?x ?t)
6	(in1 o1 ?x ?t)	←	(r-24) (out a1 ?x ?t)
0	(in1 x2 0 ?t)	←	(r-25) (obs ?t)
0	(in1 x2 1 ?t)	←	(r-26) (obs ?t)
0	(in1 a1 0 ?t)	←	(r-27) (obs ?t)
0	(in1 a1 1 ?t)	←	(r-28) (obs ?t)
0	(in2 a1 0 ?t)	←	(r-29) (obs ?t)
0	(in2 a1 1 ?t)	←	(r-30) (obs ?t)
0	(in2 x2 0 ?t)	←	(r-31) (obs ?t)
0	(in2 x2 1 ?t)	←	(r-32) (obs ?t)
0	(in2 a2 0 ?t)	←	(r-33) (obs ?t)
0	(in2 a2 1 ?t)	←	(r-34) (obs ?t)
0	(in2 o1 0 ?t)	←	(r-35) (obs ?t)
0	(in2 o1 1 ?t)	←	(r-36) (obs ?t)
0	(in1 o1 0 ?t)	←	(r-37) (obs ?t)
0	(in1 o1 1 ?t)	←	(r-38) (obs ?t)
100	falsity	←	(r-39) (in1 ?x 1 ?t) (in1 ?x 0 ?t)
100	falsity	←	(r-40) (in2 ?x 1 ?t) (in2 ?x 0 ?t)
100	falsity	←	(r-41) (out ?x 1 ?t) (out ?x 0 ?t)
100	falsity	←	(r-42) (andg ?x) (org ?x)
100	falsity	←	(r-43) (andg ?x) (xorg ?x)
100	falsity	←	(r-44) (org ?x) (xorg ?x)
100	falsity	←	(r-45) (ab0 ?x) (norm ?x)
100	falsity	←	(r-46) (ab1 ?x) (norm ?x)
100	falsity	←	(r-47) (ab0 ?x) (ab1 ?x)

Figure 11: Rules for Full Adder

Example	Gate/Connection	Fault	Correct-diagnosis ?	Time (mins)
1	x1 - x2	input floats low	Yes	1.62
2	o1	output floats high	Yes	0.48
3	x2	output floats high	Yes	0.52
4	x2	behaves as an or gate	Yes	1.24
5	a1	output floats low	No	> 20
6	x1	output floats low	No	> 20

Figure 12: Sample BRACE runs on Circuit Diagnosis

to be the same as the normality assumptions. BRACE will then consider retracting such assumptions and reexplaining them by making assumptions about what possible kinds of gates could have caused the observed outputs.

The results are summarized in the table in Figure 12. Run times are for a Sparc 1 running Lucid Common Lisp. The input consisted of sequentially giving BRACE the assumptions of inputs for 8 time instances representing all the 8 possible different sets of inputs for the circuit. After each set of inputs, the observations of the corresponding outputs for that time instance were also given to BRACE to explain.

The failure runs (examples 5 and 6) were ones where BRACE did not terminate within 20 minutes. These failures are due to inefficiencies in the backward chaining process that occur whenever large numbers of observations need to be re-explained. In the next section, we discuss this problem and our proposals for solving it.

To illustrate BRACE's diagnostic procedure, consider example 1 from Figure 12. BRACE starts off with the assumption that all gates and connections in the circuit are normal. In all our test runs, the sequence of observations presented to the system is unchanged. In this particular run, the observed outputs for the first two inputs happen to be consistent with the initial assumptions of normality of the circuit. BRACE does not change its theory and explains these observations with the given assumptions. Next, BRACE is presented with an observation where the output of x2 contradicts the predicted output. BRACE now looks for assumptions to retract in order to eliminate this contradiction. One of the attempted retractions is rule R-20 which says that the first input of x2 is the output of x1. In backward chaining on the unexplained observations (in this case, the outputs of x2 for this and the previous time instances), we look for assumptions that would explain all these observations. The first such assumption that yields a complete explanation is that of rule R-25 which says that the first input of x2 is stuck at 0. Since we now have a complete explanation, we proceed to look at the rest of the time instances and find this new theory to explain all observations presented to it. Hence the diagnosis BRACE yields is the correct one of a broken connection between gates x1 and x2 with the relevant input stuck at 0.

The trace described above shows how BRACE combines consistency-based and abductive approaches to diagnosis. The revision algorithm establishes consistency in the model of the device by retracting connection rules or normality assumptions. This is followed by the abductive step which actually explains the observed behaviour by

making assumptions of possible fault types.

We also did one experiment in order to compare the revision algorithm of BRACE to a purely abductive approach to diagnosis. We used example 1 above and modified our theory so that BRACE used pure abduction to generate diagnoses. We did not start with initial assumptions of normality but instead reduced the cost of making a normality assumption to zero i.e., normality is now considered a free assumption but it still has to be abduced. Under this scheme, the system ran for over an hour and had not yet found a complete, consistent diagnosis[9]. This result is as expected since the majority of components in most cases are normal. The purely abductive approach does not make use of this fact as BRACE does and thus has to spend a lot of time just making normality assumptions. BRACE is able to start with default assumptions of normality and then make fault assumptions only for those components where the normality assumptions had to be retracted to establish consistency. Thus our combination of abductive and consistency-based approaches gives us faster abductive diagnoses.

6 Discussion and Future Work

This section outlines several problems in the current system and our proposed approaches to solving them. We classify these into four classes of problems — improvements in the implementation, applications of our system, extensions of the current approach and analysis of our work. The first two classes of problems are ones which will definitely be included as part of the work of this proposed dissertation while the latter two are problems which we aim to study but we do not necessarily consider part of the dissertation work.

6.1 Proposed Work

6.1.1 Implementation Improvements

The current implementation of the BRACE system suffers from problems of efficiency and lack of generality that we propose to solve. As mentioned earlier, it is currently built on top of existing implementations of the ATMS (COCO) and abductive explanation (ACCEL). A number of efficiency problems arise from attempts to fit our algorithm into the framework of these two existing systems.

An example of such a problem is in the abduction routine. In order to get the ATMS to update labels correctly, our backward chaining function actually just "suggests" nodes for the forward chainer to act on. Unfortunately, running the forward chainer very often is computationally explosive since each of the newly suggested nodes is compared with all the nodes in the knowledge base with which it can make inferences.

[9] This was not the most efficient diagnostic procedure since a lot of BRACE's bookkeeping, such as the retention of all nogoods for revision, is not necessary for the abductive procedure. This incurs a heavy cost in computational time.

By moving away from the existing ATMS implementation, we hope to avoid such problems.

The current system does not unify assumptions. For instance, it may propose two assumptions (twin-star a) and (twin-star b) (where a and b are skolem constants) to explain two different observations and not even consider that a and b may be the same. This will clearly inhibit the consideration of reasonable explanations in a number of domains. The new implementation will incorporate assumption unification.

The enhancements above can be achieved by moving to Ng's latest version of ACCEL [Ng, 1992]. This version implements a very efficient backward chaining system which uses caching of partial explanations to reduce search. However, it does not include a forward chainer like earlier versions of ACCEL. Since a forward chainer is essential to BRACE, it will have to be modified to include one.

Another modification we intend to incorporate is the application of a Reiter-style algorithm [Reiter, 1987] to pick out sets for possible retraction. Currently, BRACE uses a very simple approach to generating retraction sets — picking at least one from each of the sets of minimal nogoods in order of increasing centrality. However, this does not guarantee a minimum sum of centralities because some assumptions may be found in multiple nogoods and may thus be preferable to retract even if they have higher centrality. For instance if we have two nogoods, {a c} and {b c} where a and b have centralities of 2 each and c has a centrality of 3, the retraction set with the lowest total centrality (3) is {c} and not {a b} (with a total centrality of 4) which is what BRACE would pick.

Reiter's algorithm looks at components to fault in consistency-based diagnosis by computing minimal *hitting sets* of components which would account for all the anomalies. This is analogous to choosing sets of assumptions to retract to establish consistency. By using this algorithm and picking out minimal sets with the lowest total centrality, we will be able to minimize the total centrality of the retracted sets.

6.1.2 Applications

We are looking at several domains as possible applications for BRACE. Among these are the domains of natural language plan recognition and diagnosis of time varying physical systems.

The plan recognition domain is one to which we have successfully applied BRACE's predecessor TRAIN [Subramanian and Mooney, 1991]. The problem here is to recognize an agent's actions as part of an overall plan. One example we examined in [Subramanian and Mooney, 1991] involves possible motivations for a particular visit to a hospital. For instance, a person who is known to be an employee of a particular hospital and who visits that hospital wearing a uniform is likely to be following a hospital-employment plan while someone not wearing a uniform and whose friend is sick is likely to be following a hospital-visiting plan. Abduction can be used to explain a sequence of observations by making assumptions about what plan the agent is following.

Belief revision can enhance a plan recognition system by allowing it to change the inferred plan in the face of contradictory evidence or even by changing some of the requirements of plans. For example, in the hospital example mentioned above, we might have a rule stating that hospital uniforms are white which could be contradicted by an observation of a green uniform in a particular case. The assumption that this rule holds could then be retracted to resolve this contradiction. Thus, the BRACE approach allows us not only to change plans for a particular case but also to change the knowledge base defining the requirements of plans.

We have already shown that device diagnosis is a task that is particularly suited for belief revision and abduction. One type of diagnosis we plan to address is that for continuous, dynamic systems. These problems are often modeled using systems such as QSIM [Kuipers, 1986] which qualitatively simulate behaviours of the systems. The problem is one of diagnosing faults in the model when observed behaviour is different from the behaviour predicted by QSIM. By rewriting the state successor generation of QSIM as a set of Horn-clause rules, we can convert the task into one that is suited to diagnosis using BRACE or other abductive systems.

Ng [Ng, 1992] addresses the problem of using pure abduction to diagnose such systems. Various QSIM constraints have been encoded into Horn-clause form and QSIM models for two problems — a temperature controller and water balance in the kidney — have been converted into this form. ACCEL has been used successfully in diagnosing faults in these systems. However, one problem that has arisen in trying to apply ACCEL to problems of greater complexity than these is that the beam width of the number of explanations that have to be maintained in order to get the correct diagnosis is very large. This results in very large execution times that makes ACCEL impractical on larger problems. With BRACE's ability to start with assumptions of normality and to revise beliefs, we can set beam widths lower and rely on revision to guide us toward a correct explanation.

One such larger problem that we plan to explore is the Reaction Control System for the Space Shuttle. This has been modeled using QSIM by Bert Kay [unpublished work at NASA Ames Research Center, 1991]. Diagnosis of this model is a problem that would be a practical application of abductive diagnostic systems. ACCEL, however, would not be able to compute diagnoses for this in reasonable time. We hope to get better results using BRACE.

6.2 Future Research Goals

6.2.1 Extensions

Research in Theory Revision [Richards and Mooney, 1991; Ginsberg, 1990; Ourston and Mooney, 1990; Towell *et al.*, 1990], a problem investigated in Machine Learning, has addressed a very similar problem to that which we have looked at — that of making changes to a theory to fit a set of data. Most work in Theory Revision has focussed on making certain kinds of syntactic changes to theories. These include addition and deletion of antecedents and addition of rules using induction. Theory Revision systems

make the assumption that all the faults lie in the rules of the theory and not with the atomic facts of the particular examples.

On the other hand, our work does not distinguish between rules and atomic assumptions. However, it does not utilize the syntactic structure of the rules to make theory changes. Instead, the only changes that can be made are the retraction and addition of assumptions. The only way BRACE can add a new rule is if it is already provided to it with a centrality of 0 (i.e., it is given as a possible rule but not initially believed).

These limitations of BRACE indicate that a synthesis of the two approaches may be in order. One of the efforts in our research will be to attempt to achieve this synthesis and incorporate such rule modifications into the BRACE framework. One way in which this might work would be for BRACE to propose possible syntactic variations of a rule that are compatible with the observations as possible new rules whenever a rule needs to be retracted.

Another way these ideas could be synthesized is by using rule induction to extend the theory. One of the main limitations of current belief revision theory is its inability to generalize from a set of observations. For example, suppose we are given a theory about tables that says that having four legs is one of the conditions for inferring that something is a table. We then see a number of observations of tables that only have three legs. The theory may be set up in such a way that these observations are consistent with it even though it has no theory to classify these examples as tables. The system would then just accept beliefs about various objects being tables even though nothing in the theory says they must be.

Current belief revision systems cannot generalize their theories about tables to include these instances. However, such generalizations are frequently necessary. Our research will investigate this relationship between the revision of beliefs and the refinement of theories in greater detail.

Another possible extension of our approach would be to allow observations to be retracted. There may be cases where our observations may be unreliable and retracting some observations may be preferable to attempting to re-explain them. For example, in our logic circuit example, we may discover that our tester is broken and so the readings we get for our outputs may themselves be wrong. We propose to study the effects of such retractions.

6.2.2 Analysis

The final part of our proposed work will include an attempt at a theoretical study of our model of belief revision. Peter Gärdenfors, in [Gärdenfors, 1988], proposes a formal model of revision operations on sets of beliefs. His postulates specify what constitutes a minimal semantic change in a belief set. We clearly do not follow Gärdenfors's postulates since we do not make a minimal change to the knowledge base given a new belief — abduction introduces new beliefs that are not necessary just to establish consistency.

Furthermore, the extensions we proposed in the previous section detract even more from the idea of minimal semantic change as they involve further generalization of theories. Issues such as simplicity of explanations, coherent theoretical structure etc. all contribute to what the new set of beliefs should be. We aim to look at this in greater detail and attempt a theoretical formulation of our work to define what constitute acceptable revisions in explanatory contexts.

We also propose to take a closer look at the centrality criterion and the cost metric. They are currently based respectively on numbers of observations explained and numbers of assumptions made. However, this means, for instance, that an assumption that an asteroid hit the earth is treated as having the same cost as an assumption that dinosaur fossil bones came from big creatures that once lived on the earth. Clearly, these two assumptions are on different levels of plausibility. One possible way of getting around this is to make cost of an assumption depend on the particular predicate to be assumed.

This would still not solve the problem of making coincidental sets of assumptions. Should the cost of making the assumptions that a comet fell to earth at a particular time and that an asteroid fell to earth at about the same time simply be the sum of the cost of making each of these assumptions? Coincidences of this magnitude should clearly cost more. We propose to look into ways that theories may be structured to avoid postulating coincidences except where clearly warranted by the observations. This could involve using probabilities and putting a high cost on making sets of assumptions whose combined probability is low.

The centrality criterion also poses problems of this sort. Clearly mere numbers of observations explained is not a good measure of the importance of an assumption. If some of the observations that an assumption helps explain have many alternative possible explanations, the assumption could cost less to retract than one that explains more observations. Furthermore, retracting a very central assumption does not necessarily mean that the resulting theory will have very high cost since an alternative simple explanation may be found for all the unexplained observations. The centrality criterion that we use to determine what should be retracted should more closely reflect the cost of making such retractions. This could be achieved by precomputing the cost of finding alternative explanations given a particular retraction. While this sounds on the surface to be computationally expensive, it may be justified by the gain in optimality of the resulting explanations. We aim to study the feasibility of implementing such improvements to the centrality criterion.

7 Related Research

We have already touched on some of the relationships between our research and other work in similar areas. This section summarizes some of these relationships and puts them in perspective.

Belief Revision Various belief revision systems have varying degrees of similarity to our work. Gärdenfors [Gärdenfors, 1988] is considered one of the authoritative researchers in this area. As mentioned before, his work follows a coherence theory of belief revision whereas ours is a modification on the foundational theory. While his ideas of epistemic entrenchment are related to our own idea of centrality, our work does not follow his ideas of minimal change of belief sets since we look for complete explanations of our observations rather than just minimally changed consistent sets of beliefs.

Truth Maintenance Systems Truth Maintenance Systems [Doyle, 1979; deKleer, 1987] are systems that facilitate the maintenance of consistent sets of beliefs, but do not directly incorporate abduction. The JTMS of [Doyle, 1979] incorporates belief revision using dependency-directed backtracking to establish consistency in its belief set. The JTMS, however, maintains only one set of beliefs, unlike BRACE which does a heuristic beam search on a limited number of belief sets. The JTMS also has no criterion for deciding what to retract, unlike BRACE, which uses a centrality criterion.

The ATMS [deKleer, 1987], on the other hand, maintains all possible consistent sets of beliefs while discarding inconsistent ones. This eliminates the need to revise beliefs although a lot of unnecessary computation is done to create consistent sets of assumptions that may not actually need to be computed as they may not actually be part of any overall explanatory theory.

As mentioned before, we see our algorithm as a synthesis of the JTMS and ATMS approaches in that we maintain a limited number of theories and prune our ATMS labels to get rid of environments that are not consistent with any of our current theories. At the same time, we have a JTMS-style revision algorithm to create consistent theories from inconsistent ones.

A number of researchers have extended TMS systems in various ways. One application of TMS's to abduction is [Kakas and Mancarella, 1990]. They perform abductive explanation in a logic programming framework and use Truth Maintenance to maintain consistency. However, the only kinds of revision they do are based directly on the JTMS and ATMS frameworks. Kakas & Mancarella suggest using either JTMS-style dependency-directed backtracking or ATMS-style maintenance of all consistent explanations. Furthermore, they have fixed theories that they do not question and instead focus only on choice of new assumptions.

Abduction-based Systems One application of abduction to problems of scientific discovery and associated theory changes is ABE [O'Rorke *et al.*, 1990] which was used to model changes to theories of oxygen and combustion. When ABE reaches a contradictory state, it eliminates a predetermined portion of the theory that it considers peripheral. It then uses abduction to form an alternative explanation. BRACE, on the other hand, makes no firm distinction between core and peripheral portions of the theory and only retracts parts of the theory as necessary to establish consistency.

Knowledge Assimilation The KI system [Murray, 1988] encodes a technique for assimilating rules into a knowledge base and includes a method for resolving the resulting contradictions. However, it does not build explanations using abduction like BRACE.

Theory Revision The machine learning community has taken various approaches to the problem of theory revision [Richards and Mooney, 1991; Ginsberg, 1990; Ourston and Mooney, 1990; Towell *et al.*, 1990]. These approaches have concentrated on finding fixes in the given theory that will make the data consistent with it. The (usually case-specific) data is considered accurate and the systems try to find faults in the theory. BRACE looks for faults in both the theory and the atomic data. However, we currently restrict BRACE to retraction of rules and atomic assumptions as the only theory changes considered.

8 Conclusion

We have presented a proposal for an approach to explanation that uses the paradigms of belief revision and abduction. The approach calls for maintaining a limited number of explanations and doing a heuristic beam search over these. Abduction is used to generate explanations while belief revision is used to achieve consistency when consistent explanations cannot be found. Abduction can generate new explanations in the context of these revised theories.

A preliminary implementation of our approach is the system BRACE. BRACE successfully models theory changes in the domains of scientific discovery and diagnosis of logic circuits. BRACE, however, suffers from some problems of inefficiency and lack of generality that is the focus of future research.

We have also suggested various extensions to BRACE and new and interesting applications to diagnosis of continuous dynamical systems and plan recognition. Further, we propose to look into doing a theoretical analysis of our approach and to study ways in which BRACE can be further enhanced.

Acknowledgements

The following grants have been used to support this research: IRI-9102926 from the National Science Foundation, NCC 2-629 from NASA's Ames Research Center, and Texas Advanced Research Projects grant number 003658114. I would like to thank all of these agencies for their support.

I would also like to thank Ray Mooney, my research supervisor, for his guidance throughout this project and for his very detailed comments on many earlier versions of this document; Hwee-Tou Ng and Adam Farquhar, for use of their respective imple-

mented systems; and Hwee-Tou (again) and Rich Mallory, for their comments on an earlier version of this document.

References

[Charniak and McDermott, 1985] E. Charniak and D. McDermott. *Introduction to AI.* Addison-Wesley, Reading, MA, 1985.

[Cox and Pietrzykowski, 1987] P.T. Cox and T. Pietrzykowski. General diagnosis by abductive inference. In *Proceedings of the 1987 Symposium on Logic Programming*, pages 183–189, 1987.

[Davis, 1984] R. Davis. Diagnostic reasoning based on structure and behavior. *Artificial Intelligence*, 24:347–410, 1984.

[deKleer and Williams, 1987] J. deKleer and B. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[deKleer, 1987] Johan deKleer. An Assumption-based TMS. *Artificial Intelligence*, 28:127–162, 1987.

[Doyle, 1979] J. Doyle. A truth maintenance system. *Artificial Intelligence*, 12:231–272, 1979.

[Dressler and Farquhar, 1990] O. Dressler and A. Farquhar. Putting the problem solver back in the driver's seat: Contextual control of the ATMS. In J. P. Martins and M. Reinfrank, editors, *Truth Maintenance Systems - ECAI90 Workshop*, pages 1–16, Stockholm, Sweden, August 1990. Springer-Verlag.

[Gärdenfors, 1988] Peter Gärdenfors. *Knowledge In Flux: Modeling the Dynamics of Epistemic States.* MIT Press, Cambridge, Mass., 1988.

[Ginsberg, 1990] A. Ginsberg. Theory reduction, theory revision, and retranslation. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 777–782, Detroit, MI, July 1990.

[Kakas and Mancarella, 1990] A. C. Kakas and P. Mancarella. Knowledge assimilation and abduction. In J. P. Martins and M. Reinfrank, editors, *Truth Maintenance Systems - ECAI90 Workshop*, pages 54–70, Stockholm, Sweden, August 1990. Springer-Verlag.

[Katsuno and Mendelzon, 1991] H. Katsuno and A. O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.

[Kuipers, 1986] B. J. Kuipers. Qualitative simulation. *Artificial Intelligence*, 29:289–388, 1986.

[Martins and Shapiro, 1988] J. P. Martins and S. C. Shapiro. A model for belief revision. *Artificial Intelligence*, 35:25–79, 1988.

[Muller, 1988] R. Muller. *Nemesis: The Death Star*. Weidenfeld and Nicolson, New York, New York, 1988.

[Murray, 1988] K. S. Murray. KI: An experiment in automating knowledge integration. Technical Report AI88-90, Artificial Intelligence Laboratory, University of Texas, Austin, TX, October 1988.

[Ng and Mooney, 1990] H. T. Ng and R. J. Mooney. On the role of coherence in abductive explanation. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 337–342, Boston, MA, July 1990.

[Ng and Mooney, 1991] H. T. Ng and R. J. Mooney. An efficient first-order Horn-clause abduction system based on the ATMS. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 494–499, Anaheim, CA, July 1991.

[Ng, 1992] H. T. Ng. *A General Abductive System with Application to Plan Recognition and Diagnosis*. PhD thesis, University of Texas at Austin, Austin, Texas, 1992.

[O'Rorke *et al.*, 1990] P. O'Rorke, S. Morris, and D. Schulenburg. Theory formation by abduction: A case study based on the chemical revolution. In Jeff Shrager and Pat Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufmann, 1990.

[Ourston and Mooney, 1990] D. Ourston and R. Mooney. Changing the rules: a comprehensive approach to theory refinement. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 815–820, Detroit, MI, July 1990.

[Peirce, 1958] C. S. Peirce. *Collected Papers of Charles Sanders Peirce*. MIT Press, Cambridge, Mass., 1958.

[Poole, 1989] D. Poole. Normality and faults in logic-based diagnosis. In *Proceedings of the Eleventh International Joint conference on Artificial intelligence*, pages 1304–1310, Detroit, MI, August 1989.

[Rao and Foo, 1989] A. S. Rao and N. Y. Foo. Formal theories of belief revision. In *First International Conference on Principles of Knowledge Representation and Reasoning*, pages 369–380, Toronto, Ont., 1989.

[Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

[Richards and Mooney, 1991] B. Richards and R. Mooney. First-order theory revision. In *Proceedings of the Eighth International Workshop on Machine Learning*, pages 447–451, Evanston, IL, June 1991.

[Stickel, 1988] M. E. Stickel. A Prolog-like inference system for computing minimum-cost abductive explanations in natural-language interpretation. Technical Report Technical Note 451, SRI International, Menlo Park, CA, September 1988.

[Subramanian and Mooney, 1991] S. Subramanian and R. J. Mooney. Combining abduction and theory revision. In R. S. Michalski and G. Tecuci, editors, *Proceedings of the First International Workshop on Multistrategy Learning*, pages 207–214, Harpers Ferry, WV, November 1991. Center for Artificial Intelligence, George Mason University.

[Thagard, 1992] P. Thagard. The dinosaur debate: Explanatory coherence and the problem of competing explanations. In J. Pollock and R. Cummins, editors, *Philosophy and AI: Essays at the Interface*. MIT Press/Bradford Books, Cambridge, MA, 1992.

[Towell *et al.*, 1990] G. G. Towell, J. W. Shavlik, and Michiel O. Noordewier. Refinement of approximate domain theories by knowledge-based artificial neural networks. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 861–866, Boston, MA, July 1990.

[Waters, 1990] T. Waters. The dinosaur acid test. *Discover*, February 1990.