1 Introduction

"… a new fact is a grain of sand that irritates the mind and abduction forms a pearl of wisdom around it."[1]

Abductive reasoning is a form of logical inference which aims to derive plausible explanations for a body of data. In fact, it is often regarded as "inference to the best explanation". Basically, an abductive inference works by generating or "proposing" hypotheses which would explain a group of facts (usually observations). For example, suppose we know that everyone suffering from appendicitis experiences pain in the right lower side of the abdomen; John has a pain in the right lower side of the abdomen; maybe he is suffering from appendicitis. Note that this is only one possible explanation — John may actually be suffering from some other disorder.

The term "abduction" was first introduced by the celebrated American philosopher Charles Sanders Peirce who was the first person to distinguish it as a fundamental mode of reasoning. It has been suggested [14] that the term itself is derived from the word "apogoge" which is the name of a form of syllogistic reasoning mentioned by Aristotle in *Prior Analytics*. This form of syllogism looks for premises that would make some conclusion more desirable. Since the Greek prefix "apo" means "motion away from", the word "abduction" would seem to be an appropriate translation.

1. Sowa [22] attributes this metaphor of abduction to Lewis and Mack [10].

1.1 Abduction as a Form of Logical Reasoning[2]

The notion of abduction is, in essence, easy to comprehend but is probably best understood when contrasted with the other two fundamental forms of logical inference: *deduction* and *induction* (see also [14] and [20]).

Deduction

Deductive reasoning is characterised by the following schema:

$$\begin{array}{ll} \forall x\, P(x) \rightarrow Q(x) & \text{(Major Premise)} \\ \underline{P(a)\qquad\qquad\qquad} & \text{(Minor Premise)} \\ Q(a) & \text{(Conclusion)} \end{array}$$

From a rule (major premise) and a particular case (minor premise) we can derive a new fact (conclusion). For example, suppose that all trains for the Blue Mountains depart from Platform 2 and that we know a particular train is headed for the Blue Mountains, then we can conclude that it departs from Platform 2.

Induction

Similarly, inductive reasoning can be characterised in the following manner:

$$\begin{array}{ll} P(a) & \text{(Minor Premise)} \\ \underline{Q(a)\qquad\qquad\qquad} & \text{(Conclusion)} \\ \forall x\, P(x) \rightarrow Q(x) & \text{(Major Premise)} \end{array}$$

From a particular case (minor premise) and a fact related to that case (conclusion) we hypothesise a rule (major premise) which generalises this relationship. Continuing the previous example, suppose we notice that a train is headed for the Blue Mountains and it is leaving from Platform 2, then we could hypothesise that maybe all trains for the Blue Mountains depart from Platform 2.

2. Refer to [6] for a more detailed discussion of the concepts presented in the following two sections.

Abduction

Finally, we can characterise abductive reasoning in analogous fashion:

$\forall x\ P(x) \rightarrow Q(x)$	(Major Premise)
$\underline{Q(a)}$	(Conclusion)
$P(a)$	(Minor Premise)

From a rule (major premise) and a particular fact (conclusion) we hypothesise a case (minor premise) which, if true, would explain the fact. Concluding our example, we know that all trains for the Blue Mountains depart from Platform 2 and we observe that a train is currently at Platform 2, then we could hypothesise that the train is headed for the Blue Mountains.

Peirce's main interest in logic stemmed from his desire to develop formal methods of research especially with regard to the advancement of science. His theory of abduction furnishes logic with a method for proposing hypotheses. Many philosophers do not agree that logic is involved in proposing a hypothesis but, rather, that it is only concerned with techniques for testing them.

Peirce classified the three fundamental forms of inference into two categories: *explicative inference* and *ampliative inference*. Explicative inference refers to inference where the conclusion necessarily follows from the premises while ampliative inference refers to inference where the conclusion does not necessarily follow from the premises. Therefore, deduction is a form of explicative inference while abduction and induction are forms of ampliative inference. Peirce notes that from explicative inference to ampliative inference the "security" of the inference decreases while the "productiveness" increases. Peirce not only identified the three fundamental forms of logical inference but he also maintained that they represent three stages of scientific inquiry in the following manner:

i. abduction proposes hypotheses;

ii. deduction derives the consequences of the hypotheses; and,

iii. induction tests or verifies hypotheses.

1.2 The Nature of Abduction

There are two major factors which must be considered when investigating abductive reasoning:

i. constructing or determining explanations; and,

ii. selecting the "best" or most plausible from among these.

Approaches to both of these aspects will be elaborated in the following chapters of this report. We shall briefly examine Peirce's view on the latter aspect, as he makes some interesting points which will motivate our later discussion.

Peirce suggest three major factors that are important in selecting the best hypothesis:

— a hypothesis must be capable of explaining the facts;

— we must be able to verify a hypothesis — in particular, through experiment; and,

— considerations of "economy" should guide the choice of best hypothesis.

With regard to this last factor, Peirce mentions three factors that should play a part. Firstly, it is preferable to select a hypothesis which can be verified with lesser cost with respect to "money, time, energy and thought"([6] p. 43). Secondly, we should take into account a hypothesis' effect on other projects. Finally, we need to consider the intrinsic value of a proposed hypothesis.

Peirce proposed a number of ways of ascertaining the intrinsic value of a hypothesis. The notion of "simplicity" is one important consideration. Initially, Peirce used the concept of logical simplicity to express this notion but later he maintained that the hypothesis which appears more "natural" or which is suggested by "instinct" should be considered simpler. Another consideration is the likelihood of a hypothesis, though Peirce suggests that care should be exercised when adopting measures of likelihood as most are subjective. When a hypothesis can be decomposed into parts an important consideration is to use "caution" because "twenty skillful hypotheses will ascertain what a million stupid ones will fail to do"([6] p. 50). A final consideration is the "breadth of a hypothesis". Basically, a hypothesis which can explain more facts is likely to be more useful.

It is interesting to note that, with respect to the considerations presented above, Peirce is not suggesting that the best hypothesis is "truer" but rather that it should be tested first. This hypothesis is preferred because it appears to be the most easily verified. If, ultimately, it is found to be false then we can proceed to examine another hypothesis.

1.3 Aims

Recently, in the field of artificial intelligence, there has been a lot of interest in developing a proper theoretical framework for abductive reasoning. This report aims to investigate a number of proposals for formalising various aspects of the notion of abduction. Currently, approaches can be divided into two categories: *set-cover* based and *logic* based.

In the following chapter we shall examine set-cover based approaches while in Chapter 3 we will consider various logic based approaches, examining how various approaches handle the tasks of constructing explanations and selecting the most plausible explanation. In Chapter 4, we shall present our conclusions.

2 Set-Cover Based Approaches to Abduction

"Induction is reasoning from particulars to a general law; abduction, reasoning from effects to causes ."

(Charles S. Peirce)

2.1 Introduction

The main set-based approach to abductive reasoning and the only one to be presented here is that proposed by Peng and Reggia [14], [12].

2.2 Parsimonious Covering Theory[3]

2.2.1 The Nature of a Diagnostic Problem

Peng and Reggia [14] have developed a comprehensive set-cover based approach to solving diagnostic problems through abductive inference. A diagnostic problem involves finding an explanation for the existence of a set of manifestations (observations, symptoms, effects, etc.) using existing knowledge. Peng and Reggia view the task of proposing hypotheses to form an explanation as the "resolution of two conflicting goals"([14] p. 7):

3. The definitions presented in this section have been taken from [14]. For a thorough coverage of this material, the interested reader is referred to this work.

i. Covering goal — to explain all present manifestations; and,

ii. Parsimony goal — to minimise the complexity of the explanation.

It should be evident that these two goals correspond to the two aspects of abduction mentioned in the first chapter (i.e. constructing an explanation and selecting the best explanation).

A diagnostic problem consists of two important entities: *manifestations* and *disorders*. Manifestations refer to symptoms, effects, etc. that are observable (eg. "the engine does not start", "John has a severe cough"). Disorders refer to diseases, malfunctions, etc. and are considered to be causes of manifestations (e.g. "perhaps the battery is dead", "perhaps John has a cold"). Manifestations are related to disorders by causal associations which can be exhibited through the use of a causal network (see Figure 1.).

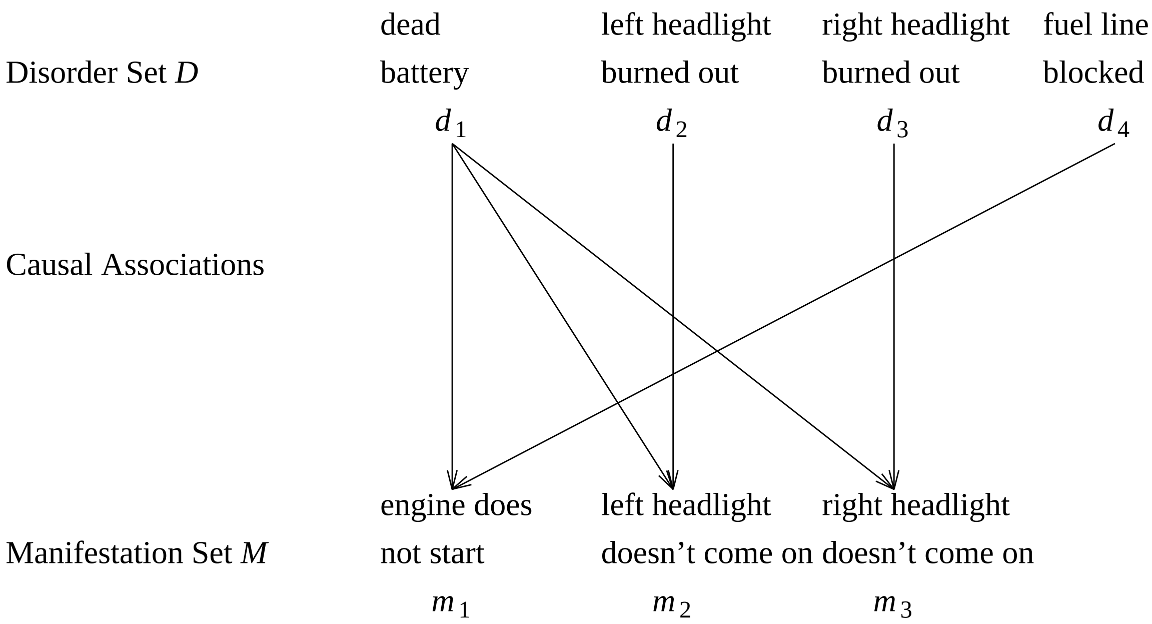


Figure 1. Causal network for a simple automotive diagnostic system.

Formally, Parsimonious Covering Theory expresses these notions through the following definition.

Definition 2.1 :

A diagnostic problem P is a 4-tuple $<D, M, C, M^+>$ where:

$D = \{d_1, \ldots, d_n\}$ is a finite, non-empty set of objects, called disorders;

$M = \{m_1, \ldots, m_k\}$ is a finite, non-empty set of objects, called manifestations;

$C \subseteq D \times M$ is a relation with $domain(C) = D$ and $range(C) = M$, called causation; and,

$M^+ \subseteq M$ is a distinguished set of M which is said to be present.

In an implementation based upon this definition, the sets D and M and the relation C would correspond to the knowledge base while the set M^+ would correspond to the input to the system. As we shall see, M^+ need not be fully specified to begin with but may be constructed incrementally.

In determining an explanation it is important to consider the effects of disorders and the causes of manifestations. Therefore, we define the following two sets:

Definition 2.2 :

For any $d_i \in D$ and $m_j \in M$ in a diagnostic problem $P = <D, M, C, M^+>$,

$effects(d_i) = \{m_j \mid <d_i, m_j> \in C\}$, the set of objects directly caused by d_i; and,

$causes(m_j) = \{d_i \mid <d_i, m_j> \in C\}$, the set of objects which can directly cause m_j.

It is also possible that more than one disorder is present. Therefore, we extend the previous definition to take account of the causes and effects of groups of items.

Definition 2.3 :

For any $D_I \subseteq D$ and $M_J \subseteq M$ in a diagnostic problem $P = <D, M, C, M^+>$,

$effects(D_I) = \bigcup_{d_i \in D_I} effects(d_i)$, and

$causes(M_I) = \bigcup_{m_j \in M_J} causes(m_j)$.

Example 2.1 :

From Figure 1. we have the following examples. (Note that we are using variables to refer to the manifestations and disorders in order to reduce the amount of space required.)

$effects(d_2) = \{m_2\}$

$causes(m_2) = \{d_1, d_2\}$

$effects(\{d_2, d_3\}) = \{m_2, m_3\}$

$causes(\{m_2, m_3\}) = \{d_1, d_2, d_3\}$

It should be quite clear that the disorders proposed as an explanation should account for the observed manifestations. To this end, we define the notion of a *cover*.

Definition 2.4 :

The set $D_I \subseteq D$ is said to be a cover of $M_J \subseteq M$ if $M_j \subseteq effects(D_I)$.

In other words, a set of disorders is considered a cover for a set of manifestations if their direct effects include all these manifestations. This allows us to define what constitutes an explanation.

Definition 2.5 :

A set $E \subseteq D$ is said to be an explanation of M^+ for a problem $P = <D, M, C, M^+>$ iff E covers M^+ and E satisfies a given parsimony criterion.

Therefore, an explanation consists of a set of disorders whose direct effects are capable of accounting for all the observed manifestations and also satisfy some selective criterion.

In order to elaborate on the nature of this parsimony criterion we shall firstly distinguish some different types of covers.

Definition 2.6 :

1. A cover D_I of M_J is said to be *minimum* if its cardinality is smallest among all covers of M_J.

2. A cover D_I of M_J is said to be *irredundant* if none of its proper subsets is also a cover of M_J; it is *redundant* otherwise.

3. A cover D_I of M^+ is said to be *relevant* if it is a subset of $causes(M^+)$;

it is *irrelevant* otherwise.

Of these, irredundant covers are usually considered the most important. Even though minimum covers explain the observed manifestations by hypothesising the presence of the least number of disorders, there are many cases where this explanation may not be the most plausible. For example, it may seem more reasonable to explain a set of symptoms by hypothesising the presence of certain common diseases rather than proposing an explanation consisting of a lesser number of relatively uncommon diseases. Relevant but redundant explanations are usually not favoured because they contain more hypotheses than are necessary to explain the manifestations present.

Example 2.2 :

Using the causal network presented in Figure 1. and $M^+ = \{m_2, m_3\}$ we have the following examples of covers:

$\{d_1\}$	relevant and minimal
$\{d_2, d_3\}$	relevant and irredundant
$\{d_1, d_2, d_3\}$	relevant and redundant
$\{d_1, d_2, d_3, d_4\}$	irrelevant

The solution to a diagnostic problem then is any cover which satisfies our selected parsimony criterion.

Definition 2.7 :

The solution of a diagnostic problem $P = \langle D, M, C, M^+ \rangle$, designated $Sol(P)$, is the set of all explanations of M^+.

It is evident that a solution may consist of many explanations. Later we shall briefly mention one method that will allow us to prune this collection.

2.2.2 An Algebra for Constructing Covers

Having defined the nature of a diagnostic problem and what constitutes a solution, Parsimonious Covering Theory provides us with an algebra for constructing a cover. This algebra is the essence of Parsimonious Covering Theory and will provide us with the necessary tools to describe an algorithm which solves diagnostic problems.

2.2

In defining this algebra, covers are represented in a very compact format known as a *generator*. A generator consists of sets of different disorders corresponding to hypotheses in a solution and can be used to generate the explanations that constitute a solution. If we have a generator $\{A, B\}$, where A and B are disjoint sets, then a hypothesis can be composed by taking one disorder from each set.

Definition 2.8 :

Let $g_1,...,g_n$ be non-empty pairwise-disjoint subsets of D. Then $G_I = \{g_1, . . . , g_n\}$ is a generator. The class generated by G_I, designated as $[G_I]$ is defined to be $[G_I] = \{\{d_1, . . . , d_n\} \mid d_i \in g_i, 1 \le i \le n\}$.

Example 2.3 :

The generator $(\{d_1, d_2\}, \{d_3, d_4\})$ [4] represents the hypotheses:

$$\{\{d_1, d_3\}, \{d_1, d_4\}, \{d_2, d_3\}, \{d_2, d_4\}\}$$

Typically, disorders grouped together within the same set in a generator account for similar manifestations. In a sense, they can be considered competing disorders.

Usually, one generator is not sufficient to represent an entire solution, so we maintain sets of generators which may be defined in the following way.

Definition 2.9 :

$[G_I] = \{G_1, . . . , G_N\}$ is a generator-set if each $G_I \in G$ is a generator and $[G_I] \cap [G_J] = \emptyset$ for $I \neq J$. The class generated by G is $[G] = \bigcup_{I=1}^{N} [G_I]$.

The first operation we introduce for this algebra is that of *division*. The purpose of division is to select from a set of disorders, representing current explanations, those that will also cover a new manifestation.

---

4. We use this more conventional notation for generators since $[G_I]$ can, in a sense, be regarded analogous to a cartesian set product except that it consists of unordered sets.

Definition 2.10 :

Let $G_I = (g_1, \ldots, g_n)$ be a generator and let $H_1 \subseteq D$ where $H_1 \neq \emptyset$. Then $Q = \{Q_k \mid Q_k \text{ is a generator}\}$ is a division of G_I by H_1 if for all k, $1 \leq k \leq n$, $Q_k = (q_{k1}, \ldots, q_{kn})$ where

$$q_{kj} = \begin{cases} g_i - H_1 & \text{if } j < k \\ g_j \cap H_1 & \text{if } j = k \\ g_j & \text{if } j > k \end{cases}$$

This definition can be also be extended to generator-sets.

Definition 2.11 :

Let G be a generator-set and $H_1 \subseteq D$ where $H_1 \neq \emptyset$. A division of G by H_1 is

$$div(G,\ H_1) = \bigcup_{G_I \in G} div(G_I,\ H_1).$$

Intuitively, the class generated by dividing G_I by H_1 (i.e. $[div(G_I, H_1)]$) contains those members of $[G_I]$ (or $[G]$) which have at least one of the disorders in H_1. The significance of division in finding solutions for diagnostic problems is that if we have a set of manifestations M_J and a set of explanations for these manifestations represented by $[G]$ (where G is a generator-set), then $div(G,\ causes(m_j))$ for a new manifestation m_j not in M_J, represents explanations of M_J which also cover m_j. Therefore, a revised set of explanations for $M_J \cup \{m_j\}$ is calculated by the division $div(G,\ causes(m_j))$. Note that, even though a generator is unordered, it can be shown that any two divisions resulting from generators with different internal orderings will be equal.

The elements of $[G_I]$ (or $[G]$) that do not cover the newly introduced manifestation can be determined by an operation known as the *residual* of division.

Definition 2.12 :

Let $G_I = (g_1, \ldots, g_n)$ be a generator, G a generator-set, and $H_1 \subseteq D$ where $H_1 \neq \emptyset$. Then the residual of a division of G_I by H_1 is

$$res(G_I,\ H_1) = \begin{cases} \{(g_1 - H_1, \ldots, g_n - H_1)\} & \text{if } g_i - H_1 \neq \emptyset \text{ for all } i, 1 \leq i \leq n, \\ \emptyset & \text{otherwise;} \end{cases}$$

and the residual of a division of G by H_1 is

$$res(G,\ H_1) = \bigcup_{G_I \in G} res(G_I,\ H_1).$$

The residual operation works by removing all elements of H_1 from G_I. This ensures that the generator-set produced will not generate something with an element of H_1. It can be shown that the division and residual operations will partition the class generated by a particular generator (or generator-set).

We can generalise the operations of division and residual so that instead of using sets of disorders, H_1, we can use generators that represent collections of disorders. So, for instance, division of G_I by a generator Q_J represents those elements of $[G_I]$ containing at least one element from each set of disorders within Q_J. The residual also has an analogous meaning and we have the following definitions.

Definition 2.13 :

Let G and Q be generator-sets, $G_I \in G$ and $Q_J \in Q$ be generators, and $q_j \in Q_J$. Then a division of G_I by Q_j is

$$div(G_{I,}\ Q_J) = \begin{cases} \{G_I\} & \text{if } Q_J = \emptyset, \\ div(div(G_I,\ q_j),\ Q_J - (q_j)) & \text{otherwise;} \end{cases}$$

a division of G by Q_J is

$$div(G,\ Q_J) = \bigcup_{G_I \in G} div(G_I,\ Q_J).$$

Definition 2.14 :

Let G and Q be generator-sets, $G_I \in G$ and $Q_J \in Q$ be generators, $q_j \in Q_J$. Then a residual of division of G_I by Q_J is

$$res(G_I,\ Q_J) = \begin{cases} \emptyset & \text{if } Q_j = \emptyset, \\ res(G_I,\ q_j) \cup res(div(G_I,\ q_j),\ Q_J - (q_j)) & \text{otherwise;} \end{cases}$$

a residual of division of G by Q_J is

$$res(G,\ Q_J) = \bigcup_{G_I \in G} res(G_I,\ Q_J);$$

and a residual of division of G by Q is

$$res(G,\ Q) = \begin{cases} G & \text{if } Q = \emptyset, \\ res(res(G,\ Q_J),\ Q - \{Q_J\}) & \text{otherwise.} \end{cases}$$

Thus far, we have seen the division and residual operations that partition the current collection of explanations into those that cover both M^+ and m_j and those that only cover M^+. Notice, in particular, that if we divide a generator G_I, the resulting generator will represent covers of the same cardinality as those represented by G_I. It is possible though that covers of higher cardinality may also constitute irredundant covers. Therefore, we define one last operation, the *augmented residual*, that adds disorders from $causes(m_j)$ to the results of a residual in order to cover both M^+ and m_j. By augmenting the residual $res(G_I, causes(m_j))$ by choosing disorders in $causes(m_j)$ not in G_I we avoid producing duplicate covers.

Definition 2.15 :

Let $G_I = (g_1, \ldots, g_n)$ be a generator, G a generator-set, and $H_1 \subseteq D$ where $H_1 \neq \emptyset$. Then the augmented residual of division G_I by H_1 is

$$augres(G_I, H_1) = \begin{cases} \{(g_1 - H_1, \ldots, g_n - H_1, A)\} & \text{if } g_i - H_1 \neq \emptyset, 1 \leq i \leq n,\ A \neq \emptyset, \\ \emptyset & \text{otherwise;} \end{cases}$$

where $A = H_1 - \bigcup_{i=1}^{n} g_i$.

The augmented residual of division G by H_1 is

$$augres(G, H_1) = \bigcup_{G_I \in G} augres(G_I, H_1).$$

Thus, from the set of explanations for M^+ and from the disorders in $causes(m_j)$ we can utilise the division and augmented residual operations to determine all explanations of $M^+ \bigcup \{m_j\}$.

Example 2.4 :

The following examples examine the division, residual and augmented residual operations.

$div((\{d_1, d_4\}), \{d_1, d_2\}) = \{(\{d_1\})\}$

$augres((\{d_1, d_4\}), \{d_1, d_2\}) = \{(\{d_4\}, \{d_2\})\}$

$div((\{d_1, d_2\}\{d_7, d_8, d_9\}), \{d_2, d_3, d_5, d_6\}) = \{(\{d_2\}\{d_7, d_8, d_9\})\}$

$res((\{d_1, d_2\}\{d_7, d_8, d_9\}), \{d_2, d_3, d_5, d_6\}) = \{(\{d_1\}\{d_7, d_8, d_9\})\}$

$augres((\{d_1, d_2\}\{d_7, d_8, d_9\}), \{d_2, d_3, d_5, d_6\}) = \{(\{d_1\}\{d_7, d_8, d_9\}\{d_3, d_5, d_6\})\}$

Armed with the three operations presented above we are now in a position to present an algorithm for solving diagnostic problems.

function BIPARTITE(D,M,C)
  variables
    m_{new} **manifestation**,
    H_1 **set of manifestations**,
    hypothesis **generator-set**;
  begin
    hypothesis = {∅}; {* initially no manifestations present *}
    while *Moremanifs* **do** {* while another manifestation exists *}
      m_{new} := *Nextman*; {* obtain the new manifestation *}
      $hypothesis := revise(hypothesis,\ causes(m_{new}))$
        {* update *hypothesis* by $causes(m_{new})$ *}
    endwhile;
    return *hypothesis* {* generator-set for $Sol(P)$ *}
  end.

The algorithm is provided with a set of disorders (*D*), a set of manifestations (*M*) and a knowledge-base (*C*) containing the causal relationships between *D* and *M*. It maintains a generator-set (*hypothesis*) representing irredundant covers for the current set of manifestations. Manifestations are introduced one at a time (m_{new}) in such a way that the generator-set *hypothesis* is modified by function *revise* so that it will then represent the irredundant covers of m_{new} and the other known manifestations. Function *revise* is in fact the heart of the algorithm and may be defined by the following formula:

$$revise(G,\ H_1) = F \cup res(Q,\ F)$$
$$\text{where } F = div(G,\ H_1) \text{ and } Q = augres(G,\ H_1)$$

The residual $res(Q,\ F)$ is used to remove any duplicate or redundant covers that may arise from the augmented residual. The significance of the rest of the formula should be straightforward from the foregoing discussion. Once all present manifestations have been dealt with, the generator-set *hypothesis* is returned and represents the solution to the diagnostic problem. The order in which the present manifestations are processed will, at most, affect the appearance of the generator-set *hypothesis* but the final solution will always be the same.

Example 2.5 :

For $M^+ = \{m_1, m_2, m_3\}$ and the causal diagram in Figure 1. the following table traces the development of the generator-set *hypothesis* using the above algorithm.

m_{new}	$causes\,(m_{new})$	*hypothesis*
m_1	$\{d_1, d_4\}$	$\{(\{d_1, d_4\})\}$
m_2	$\{d_1, d_2\}$	$\{(\{d_1\}), (\{d_2\}, \{d_4\})\}$
m_3	$\{d_1, d_3\}$	$\{(\{d_1\}), (\{d_2\}, \{d_3\}, \{d_4\})\}$

2.2.3 Extensions

So far, we have investigated a method for obtaining irredundant covers for a set of observed manifestations. Peng and Reggia also introduce a probabilistic method for determining the best explanation.

Most probabilistic approaches to problem-solving use a variant of Bayes' rule to calculate the posterior probability of some situation. Bayes' rule tells us that the posterior probability of a disorder given the manifestations is determined by the prior probabilities of the disorder and manifestations and the conditional probability of the manifestations given the disorder:

$$P(d_i \mid M_J) = \frac{P(M_J \mid d_i)\,P(d_i)}{P(M_J)}$$

Clearly, the number of probabilities required for such calculations in a reasonably sized diagnostic problem is likely to be very large.

Peng and Reggia attempt to reduce the number of probabilities needed by introducing the notion of a *causation event* — the event that d_i actually causes m_j and denoted $m_j{:}d_i$. Causation events are used to define a *conditional causal probability* $P(m_j{:}d_i \mid d_i)$ — (i.e. $P(d_i$ *actually causes* $m_j \mid d_i)$ — and use this in place of the ordinary conditional probability in the above formula. This allows us to assign a probability of 0 to situations where d_i cannot cause m_j and thus will reduce the number of probabilities that are required. Note that we can determine whether a disorder d_i can actually cause a manifestation m_j by inspecting the causal network for a given diagnostic problem and therefore we can, in a sense, consider the conditional causal probability $P(m_j{:}d_i \mid d_i)$ a numeric value associated with each link $<d_i, m_j> \in C$ of a causal network.

Thagard [24] agrees that the notion of conditional causal probability captures the intuitive notion that abductive inference is concerned with whether a disorder really explains a symptom rather than whether the two are probabilistically related but points out that it is difficult to attach an interpretation to this measure. He also disagrees with some of the independence assumptions adopted by Peng and Reggia.

In this investigation of Parsimonious Covering Theory we have only looked at the simplest kind of diagnostic problems — those where the causal network is bipartite (i.e. only two layers of entities are involved: disorders and manifestations). Many diagnostic problems also consist of intermediate layers known as *pathological states*. These entities are not disorders nor are they directly observable like manifestations but, rather, they represent abstractions. For example, the manifestation "severe pain in upper left side of chest" may be caused by the pathological state "heart attack" which may be caused by the disorder "aneurism". Many of the concepts presented in this chapter are extended to this more general form of diagnostic problem and several algorithms are developed for solving such problems. We shall not cover this material here but, instead, refer the interested reader to [14].

One last extension proposed by Peng and Reggia is a connectionist model for solving diagnostic problems. In this case, the search for a solution is treated as an optimisation problem. Again, the interested reader is referred to [14].

2.3 Conclusions

Parsimonious Covering Theory provides a well thought out, practical approach to diagnostic problem solving. It permits the confirmation of multiple disorders in a very natural way. Many traditional rule-based and statistical pattern classification systems are incapable of doing this except with the aid of some simple heuristics. Another advantage is that it allows us to present observed manifestations in an incremental manner. This means that we do not have to recalculate a solution from the beginning if a new manifestation is observed. The Parsimonious Covering Theory framework also facilitates methods for justifying explanations and generating questions for guiding a diagnostic task.

One disadvantage with Parsimonious Covering Theory is that it relies heavily on the causal diagram for representing problems. In its current state, this means that it cannot deal with negative evidence([24] — the absence of a manifestation would rule out a disorder as a possible cause). Also, Parsimonious Covering Theory only finds explanations for conjunctions of manifestations and cannot deal with more general situations.

A number of problems with Parsimonious Covering Theory have been raised by Thagard [24]. Firstly, he mentions that the notion of parsimony in terms of irredundant covers may not provide the best explanation in all cases. In some cases, it may be possible that the presence of more than one disorder account for a manifestation. For example, suppose someone has a cough and a runny nose, then it may be possible that they are suffering from both a cold and hay fever. Another problem occurs when a cover cannot be found for the observed manifestations[5] — in fact, in many cases an acceptable explanation need not cover all observed manifestations. For example, Newton's particle theory of light was an acceptable theory even though it did not account for every known behaviour of light. In such cases, Thagard suggests that the property of minimum cardinality has the advantage of being applicable even though a cover does not exist. The fact that every manifestation must be caused by some disorder also means that the system cannot discover disorders to explain the manifestations. This does not comply with Peirce's idea that abductive inference is concerned with discovery.

Allemang et. al. [1] show that the task of finding an explanation (in terms of a minimum cover) for a diagnostic problem using this set-based approach is NP-complete. They also point out some assumptions inherent in this approach. Firstly, it

5. Note that by the definition of a diagnostic problem provided by Parsimonious Covering Theory (Definition 2.1) this situation is in fact impossible.

assumes that the *explanatory power* (i.e. the set of manifestations for which a disorder can account — $effects(d_i)$) of a set of hypotheses is computable. This is usually a reasonable assumption as the causal structure of a problem is often encoded in a suitable form (e.g. in a causal network). This approach also assumes that the following *independence assumption* holds:

If S and T are subsets of D, then $effects(S \cup T) = effects(S) \cup effects(T)$.

This assumption basically says that if we add disorders, then we shall have at least the original set of manifestations and possibly more. Most diagnostic domains obey this assumption since devices can be decomposed into independent components and hypotheses concerning these components are usually independent. It is perceivable that this assumption may not hold in certain situations (e.g. two reactions may separately be accounted for by the presence of certain distinct objects but the occurrence of both reactions together may be accounted for by a totally different object).

3 Logic Based Approaches to Abduction

"each chief step in science has been a lesson in logic"
(Charles S. Peirce)

3.1 Introduction

There are a number of logic based approaches to abductive reasoning. Logic based approaches have the advantage of being more general than set-cover based approaches. In this chapter, we shall firstly investigate logic based methods for constructing explanations and then discuss approaches to determining the best explanation which can be applied to these methods. In some cases, where the method of constructing an explanation has an associated method for determining the best explanation, these two approaches will be presented together.

3.2 Constructing an Explanation

3.2.1 Mechanisation of Abduction

For historical reasons we shall begin this chapter with an investigation of how to mechanise the process of abduction. Pople [20] uses a modified resolution procedure as the basis for a method of performing abductive inference. We shall briefly describe this deduction procedure which attempts a direct proof of the conjecture from the axiom set rather than the normal resolution strategy of adding the negation of the conjecture to the axiom set and attempting to derive the null clause.

Firstly, the axiom set A is converted to conjunctive normal quantifier-free form (i.e. clausal form) and the conjecture C, which we are attempting to prove, is converted into a disjunctive normal form in which universally quantified variables are replaced by skolem functions whose arguments consist of variables from any existential quantifiers before the universal quantifier and where existential quantifiers are simply eliminated. We note that the conjecture now consists of n disjuncts $(D_1, \ldots, D_n)$ which can be converted into a series of expressions of the form $(\neg B_i \rightarrow D_i)$ where B_i is the disjunction of all D_j for $j \neq i$. $\neg B_i$ can now be added to the axiom set A and the problem becomes the derivation of D_i. This process can be performed for each disjunct in the conjecture and will result in alternative expressions of the problem.

A similar process can be performed on the axiom set. Any unit clauses will be considered facts but other clauses can be converted into a conditional form. A clause $\{L_1, \ldots, L_n\}$ can be considered an implication of the form $(M \rightarrow L_i)$, where M consists of the conjunction of all $\neg L_j$ for $j \neq i$. This procedure is performed for each literal L_i resulting in production rules. A rule of the form $(L_i \leftarrow M)$ has the intuitive meaning that attempting to prove L_i can be replaced by the problem of proving the literals that comprise M.

A conjecture C is now proved by taking one formulation of the problem as described previously and attempting to derive each of the conjuncts (i.e. attempt to prove each conjunct in one of the D_i's). This is done by keeping a list of conjuncts and, taking each conjunct in turn, attempting to match it with the left-hand-side of the production rules (i.e. part to the left of the '$\leftarrow$'). If a match is made with a unit axiom then a "solution" has been found and the term may be deleted from the list after instantiating the appropriate variables. If a match is made with the left-hand-side of a rule then the term is removed from the list and replaced by the terms on the right-hand-side of the rule after the appropriate variables have been instantiated. Sometimes a subgoal is generated which is the negation of what we are trying to prove. In this case, as long as the other conjunctive goals can be proved, we can consider this a proof of the original goal by reductio ad absurdum. This process continues until the list of terms is empty. If this is not possible, a proof cannot be found for the conjecture using the given axiom set.

Having described the deductive framework in which we shall work, we shall now elaborate on how abductive inference is performed within this framework. If we are attempting to explain a statement and we can provide a proof for it in the above manner, then the proof tree would constitute an explanation. Abductive inference though will provide an explanation whether a proof exists for a statement or not. In this case, when attempting to prove a statement, we shall firstly proceed with our deduction procedure. When we come across a term that does not match any rule then we basically treat that term as a possible hypothesis in our explanation. A possible hypothesis is seriously considered if it is generated by at least two partial search trees during our proof (this is easily determined by observing whether the term appears two or more times in the list of terms that are to be proved). This criteria expresses the notion that a hypothesis is "better" if it explains more and is basically an implementation of the principle of Occam's Razor. This process is known as *synthesis* and is the only way in which abductive inference is performed within this framework.

The following example should make this procedure clearer.

Example 3.1 :

Suppose we have the following axiom set for a simplified medical diagnostic problem:

A1: {chills ← presence(P, S) ∧ inflammatory(P)}
A2: {pain(R) ← presence(P, S) ∧ located–in(S, R)}
A3: {inflammatory(abscess)}
A4: {located–in(liver, right–upper–quadrant)}
A5: {jaundice ← presence(P, liver)}

We observe the following symptoms:

{chills, pain(right–upper–quadrant}

which we must explain. A solution to this problem will be a hypothesis of the form

{presence(P, S)}

(where P is some abnormal process and S some structure).

The following illustrates how abductive inference is performed according to the procedure described above. Hypotheses are placed to the left of the vertical bar and the process is complete when there are no terms remaining to the right of the vertical bar.

1.	{| chills, pain(right–upper–quadrant}	observation
2.	{| presence(X, Y) , located–in(Y, right–upper–quadrant) , chills}	1, A2
3.	{| presence(U, V) , inflammatory(U) , presence(X, Y) , located–in(Y, right–upper–quadrant) }	2, A1
4.	{presence(U, V) | inflammatory(U) , located–in(V, right–upper–quadrant) }	3, synthesis
5.	{presence(U, liver) | inflammatory(U) }	4, A4
6.	{presence(abscess, liver) |}	5, A3

The abductive inference procedure described in this section is quite straightforward and should be capable of finding explanations for quite general statements. This procedure, though, does not tackle the task of generating alternative explanations. Pople actually suggests that the explanation generated can be tested and any new observations can then be added to the system and subjected to the abductive procedure described. Also, it implements a rather crude method for selecting the best hypothesis.

A similar method to the one presented above is presented by Cox and Pietrzykowski [4]. They proceed in a more traditional resolution style by converting the knowledge base K and an effect e, which is to be explained, into clausal form and then perform linear resolution using each element of $\neg e$. In order to ensure that an explanation does not degenerate to the deductive case, they assume that K does not imply e and, therefore, a refutation will not result from the resolution procedure. Actually, the resolution procedure will either not terminate or terminate in a *dead end* clause d. Once the set D of dead end clauses has been determined, we then negate each $d \in D$ to obtain a conjunction $p_1 \wedge \cdots \wedge p_n$, where each p_i is a literal. This expression can then be converted, through reverse Skolemisation, into a closed literal of the form $Q_1 x_1 \cdots Q_k x_k \wedge(q_1, \ldots, q_n)$, where $Q_1, \ldots, Q_k$ represent quantifiers and the $q_1, \ldots, q_n$ are derived by replacing Skolem terms in $p_1, \ldots, p_n$ with universally quantified variables. This final expression represents the explanation (cause) of the effect e and it can be shown that no other element of K is a cause of it.

3.2

3.2.2 Poole's Default Logic

Poole [17], [19], [15], [16], [18] presents a logical framework, for performing default reasoning, which is abductive in nature. His framework is developed in terms of the first-order predicate calculus and he argues that there is no need to define a new logic but rather, the way logic is used should be changed. In fact, Poole claims that his method subsumes the "intuition behind Reiter's default logic"([15] p. 27).

The main idea behind this framework is that we require a set of hypotheses which, together with a set of facts, entail some goal. In order to achieve this, the following three types of formulae are used:

$\mathcal{F}$ set of facts;

Δ set of possible hypotheses; and,

g observation to be explained.

The set of facts are formulae representing those things we know to be true in the world we are attempting to represent and that we are not prepared to give up. We assume these formulae are consistent. The set of possible hypotheses are things we are prepared to accept in constructing an explanation. We can use any ground instances of these formulae as hypotheses as long as they are consistent. By thinking of these formulae as possible hypotheses the abductive nature of this formalism should become clear. On the other hand, if we think of these formulae as defaults then we get default reasoning.

We firstly define the notion of a *scenario* which represents a set of hypotheses and facts which could be true with respect to what we know.

Definition 3.1 :

A scenario of $\mathcal{F}, \Delta$ is a set $D \cup \mathcal{F}$ where D is a set of ground instances of elements of Δ such that $D \cup \mathcal{F}$ is consistent.

This allows us to introduce a notion of explanation.

Definition 3.2 :

If g is a closed formula, then an explanation of g from $\mathcal{F}, \Delta$ is a scenario of $\mathcal{F}, \Delta$ which implies g.

So, an explanation is basically a scenario which implies our observation. We can express this another way by saying that g is explainable from $\mathcal{F}, \Delta$ if there is a set D

of ground instances of elements of Δ such that

i. $\mathcal{F} \cup D \models g$, and

ii. $\mathcal{F} \cup D$ is consistent.

The scenario $\mathcal{F} \cup D$ can then be considered an explanation of g. These two conditions say that D predicts g and that D does not predict something known to be false.

Having determined a set of hypotheses that will explain the observations, it would be interesting to see what other things are implied by these hypotheses.

Definition 3.3 :

An extension of $\mathcal{F}, \Delta$ is the set of logical consequences of a maximal (with respect to set inclusion) scenario $\mathcal{F}, \Delta$.

Example 3.2 :

Suppose we have the following set of facts

$$\begin{aligned} \mathcal{F} = \{ & \forall x\ emu(x) \Rightarrow bird(x), \\ & \forall x\ emu(x) \Rightarrow \neg flies(x), \\ & emu(polly), \\ & bird(tweety) \}. \end{aligned}$$

and the following set of possible hypotheses

$$\Delta = \{ bird(x) \Rightarrow flies(x) \}$$

Then we could explain $flies(tweety)$ using the consistent hypothesis

$$\{ bird(tweety) \Rightarrow flies(tweety) \}.$$

We cannot explain $flies(polly)$ because the hypothesis

$$\{ bird(polly) \Rightarrow flies(polly) \}$$

is inconsistent with the facts (i.e. $\neg flies(polly)$ is explainable from the facts above).

Poole shows that this formalism is compact (i.e. any g that can be explained, can be done so with a finite scenario) and that it possesses a *monotonicity* property in that anything that can be explained with a set of facts and a set of hypotheses can also be explained using the same set of facts and a superset of the hypotheses (if the set of facts change then monotonicity is not guaranteed). The more important result though is the following.

Theorem:

α is explainable if and only if α is in some extension.

Poole also introduces a method for naming defaults which simplifies both the writing of the notation and the implementation of a system based on this formalism. If we have a default $w[\mathbf{x}]$, where $\mathbf{x} = x_1, \ldots, x_n$ are free variables, then it can be named by $p_w(x_1, \ldots, x_n)$ where p_w is an n-ary predicate symbol that does not appear in the set of facts $\mathcal{F}$ or possible hypotheses Δ. The definitions of $\mathcal{F}$ and Δ now become:

$$\Delta' = \{p_w(\mathbf{x}) \text{ such that } w[\mathbf{x}] \in \Delta\},$$

$$\mathcal{F}' = \mathcal{F} \cup \{\forall \mathbf{x}\, p_w(\mathbf{x}) \Rightarrow w[\mathbf{x}] \text{ such that } w[\mathbf{x}] \in \Delta\}.$$

Poole [15] shows that the ability to name defaults does not change the semantics of the formalism in any way.

In certain situations we may wish to rule out the applicability of some hypotheses. This can be done through the use of *constraints*. Constraints are formulae which are used to reject scenarios since scenarios must be consistent with them but they cannot be used as part of an explanation.

Definition 3.4 :

A scenario of $\mathcal{F}$, C, Δ is a set $D \cup \mathcal{F}$ where D is a set of ground instances of elements of Δ such that $D \cup \mathcal{F} \cup C$ is consistent.

The previous definitions and theorems can be extended in the obvious way.

Example 3.3 :

This example shows the use of named defaults and constraints.

$$\mathcal{F} = \{bird(polly), \neg flies(bruce)\}$$

$$\Delta = \{birdsfly(x) : flies(x) \Leftarrow bird(x)\}$$

$$C = \{\neg birdsfly(x) \Leftarrow \neg flies(x)\}$$

We can explain $flies(polly)$ through the hypothesis

$$\{birdsfly(polly) : flies(poll) \Leftarrow bird(x)\}.$$

Normally we would be able to explain $\neg bird(bruce)$ through the use of the contrapositive of the default rule $birdsfly(bruce)$ but the constraint now precludes use of this rule. Finally, we would usually have been able to prove $flies(david)$ (or $flies(a)$ for any constant symbol a not in the set of facts) but, again, the constraint now prevents us from doing so.

3.2

In an earlier version of this formalism, Poole [17] defines a notion of "more general solution" which is used to select the preferred explanations. If two solutions (theories) are comparable, then the "more specific solution" is considered to be more applicable and therefore preferable. In one of his latest papers on this formalism, Poole [18] introduces a probabilistic approach based on Bayes’ rule for determining the best explanation. As with other probabilistic approaches, he makes a number of assumptions to reduce the complexity of the calculations involved.

The formalism described here has been implemented in the Theorist system [19]. One of the problems with this approach, which is also inherent in many of the other formalisms presented here, is that, since we can only select hypotheses that come from a certain set, discovery is ruled out.

An approach that is similar to that of Poole’s is also proposed by Eshghi and Kowalski [5]. They define an *abductive framework* and an *abductive solution* in the following way

Definition 3.5 :

$\langle T, I, A \rangle$ is an abduction framework iff

T is a Horn clause theory (without denials),

I is a set of integrity constraints,

A is a set of predicate symbols, called *abducible* predicates.

Given the abduction framework $\langle T, I, A \rangle$, the hypothesis set Δ is an abductive solution for the existentially quantified conjunction of atoms Q iff

Δ is a set of variable free abducible atoms,

$T \cup \Delta \models Q$

$T \cup \Delta$ satisfies I.

It is then shown that this formulation of abduction is in fact a generalisation of negation by failure.

3.2.3 Konolige's Abductive Frame

Konolige [8] has attempted to present a general theory of abduction that formalises some of the notions developed in other approaches. An abduction problem is stated with respect to an *abductive frame* which is formally defined in the following way.

Definition 3.6 :

An abductive frame is a tuple $<\mathcal{L}, \Sigma, \mathcal{A}, \mathcal{O}>$, where

$\mathcal{L}$ is a logical system;

Σ is a set of sentences of $\mathcal{L}$, the background theory;

$\mathcal{A}$ is the assumption vocabulary; and,

$\mathcal{O}$ is the observation vocabulary.

Here, $\mathcal{L}$ is an arbitrary logical system with a well defined notion of consequence (written $\Gamma \vdash_{\mathcal{L}} \phi$). The background theory Σ contains knowledge of causal relations within the world while the observation vocabulary $\mathcal{O}$ expresses what is observed in the world and the assumption vocabulary $\mathcal{A}$ represents what we may assume to explain the observations.

Given an abductive frame, we can now define what constitutes an explanation.

Definition 3.7 :

Let $<\mathcal{L}, \Sigma, \mathcal{A}, \mathcal{O}>$ be an abductive frame. Let S, the situational facts, be a finite set of sentences not containing the vocabulary of $\mathcal{O}$. Let O, the observations, be a sentence from the vocabulary of $\mathcal{O}$.

An explanation of the observations is a finite set $A \subset \mathcal{A}$ such that

1. $\Sigma \cup S \cup A \vdash_{\mathcal{L}} O$.

2. A is consistent with Σ.

3. $\mathcal{A}$ is minimal.

A cautious explanation is the disjunction of all explanations, that is, $\vee_i A_i$.

The situational facts, correspond to knowledge of a particular situation that does not have to be explained. It is not included in the background theory Σ since abductive frames are intended to be applicable in many situations. The minimality condition on A expresses only the condition that there is no proper subset of A which is also an explanation. Therefore, there may be more than one candidate for A, but the question of distinguishing between them is not considered further as it may depend on the particular domain.

Example 3.4 :

Using default logic as our logical system and the following sets:

$$
\begin{aligned}
\Sigma = \{ & \forall t\ rain(t) \supset wet-lawn(suc(t)) \\
& \forall t\ rain(t) \supset wet-road(suc(t)) \\
& \forall t\ sprinkler(t) \supset wet-lawn(suc(t)) \\
& \forall t\ sun(t) \supset dry-road(suc(t)) \\
& wet-lawn(t) : wet-lawn(suc(t)) / wet-lawn(suc(t)) \\
& dry-lawn(t) : dry-lawn(suc(t)) / dry-lawn(suc(t)) \\
& wet-road(t) : wet-road(suc(t)) / wet-road(suc(t)) \\
& dry-road(t) : dry-road(suc(t)) / dry-road(suc(t)) \\
& \forall t\ dry-lawn(t) \equiv \neg wet-lawn(t) \\
& \forall t\ dry-road(t) \equiv \neg wet-road(t) \}
\end{aligned}
$$

$$S = \{rain(1),\ dry-road(0),\ suc(0) = 1,\ suc(1) = 2\}$$

$$O = \{dry-lawn(1)\}$$

The assumption vocabulary $\mathcal{A}$ will contain properties of the initial situation (e.g. $\{rain(0),\ sprinkler(0),\ wet-lawn(0),\ dry-lawn(0)$, etc.$\}$ as these represent causes of future situations. With the sets above, the only explanation is

$$dry-lawn(0).$$

This can be used to conclude $dry-lawn(1)$, $dry-road(1)$, $wet-road(2)$ and $wet-lawn(2)$ and the corresponding negative literals.

This theory is quite general since the logical system and background theory are not confined to a particular logic.

3.2.4 Different Forms of Abduction

Stickel [23] has distinguished a number of different forms of abduction which are related to the *level of specificity* of an explanation and he claims that they are suited to different types of applications. He considers explanations of conjunctions of positive literals from a Horn clause knowledge base but his ideas are applicable to other formulations of abduction. Under this system, an explanation for an expression $Q_1 \wedge \cdots \wedge Q_n$ is considered to be a substitution θ and a set of assumptions $P_1, \ldots, P_m$ such that $(P_1 \wedge \cdots \wedge P_m) \rightarrow (Q_1 \wedge \cdots \wedge Q_n)\theta$ is a consequence of the knowledge base.

3.2.4.1 Most Specific Abduction

The first form of abduction is know as *most specific abduction*. In this case, most specific explanations are preferred by allowing only pure literals (i.e. literals that cannot be resolved with any clauses contained in the knowledge base) to be assumed. We have already seen this form of abduction in the procedures of Pople [20] and Cox and Pietrzykowski [4].

Stickel claims that this form of abduction is suited to diagnostic tasks as it tends to neglect trivial explanations. For example, suppose we are trying to explain "the computer is malfunctioning" then suggesting the trivial explanation, that the computer is malfunctioning, is unlikely to help whereas suggesting that one of the circuit boards is faulty is likely to be of more use. Of course, sometimes explanations can be too specific. For instance, in our last example an explanation in terms of specific components or sub-components of the circuit board may be of less use than simply hypothesising that the circuit board is at fault.

Most specific abduction also makes the implicit assumption that the knowledge base is complete with respect to causes, subtypes, etc. For example, Suppose we have the following axioms in our knowledge base:

$$\mathrm{mercury}(x) \rightarrow \mathrm{liquid}(x)$$
$$\mathrm{water}(x) \rightarrow \mathrm{liquid}(x).$$

If we are attempting to explain $\mathrm{liquid}(a)$ then we would do so in terms of $\mathrm{mercury}(a)$ or $\mathrm{water}(a)$. But perhaps a is in fact something not in our knowledge base, such as milk.

Another problem with most specific abduction is that it may not compute every reasonable explanation. For example, suppose we have a knowledge base containing $P(a)$ and $Q(b)$ and we are trying to explain instances of the expression $P(x) \wedge Q(x)$. We could explain this formula by assuming $P(b)$ or $Q(a)$ but not by assuming $P(x)$ or $Q(x)$. Also, we would not be able to hypothesise $P(c)$ and $Q(c)$ (or any other value of x except a or b) to explain the expression using this form of abduction. As Stickel points out, "most specific abduction does not ‘lift’ properly from the case of ground (variable free) formulas to the general case".

3.2.4.2 Predicate Specific Abduction

The second form of abduction is called *predicate specific abduction*. This form of abduction constructs explanations by allowing only a subset of predicates to be used as hypotheses. In a sense, this notion is similar to that of Poole’s default logic. These predicates are known as assumable predicates. Stickel claims that this form of abduction is suited to planning and design-synthesis problems.

For example, if a knowledge base consists of $P(a)$ and $Q(b)$ then we could explain $P(x) \wedge Q(x)$ by: hypothesising $P(b)$ if P is an assumable predicate; hypothesising $Q(a)$ if Q is an assumable predicate, and; hypothesising $P(x) \wedge Q(x)$ if P and Q are both assumable predicates. We can see from this example that predicate specific abduction avoids some problems associated with most specific abduction. A problem with this approach is that it may be difficult to determine which predicates should be assumable. Also, it may be the case that some predicates should be assumable in some situations and not in others.

3.2.4.3 Least Specific Abduction

The final form of abduction introduced by Stickel is known as *least specific abduction*. In this case, it is argued that the most trivial explanation is the one we desire. This form of abduction is best suited to natural-language interpretation. Returning to one of our previous examples, if we are trying to explain "the computer is malfunctioning" then, in this case, we will just hypothesise that the computer has a malfunction. It would be needless to search for a more specific explanation as we may be uninterested in such matters. The implication is that when we interpret a sentence, the meaning is near the surface.

This approach has the advantage over most specific abduction in that it does not attempt to provide very specific explanations which often involve riskier assumptions. In the liquid example presented earlier, we could explain liquid(a) by hypothesising that a is a liquid rather than being forced to make wrong assumptions. In the situation where the knowledge base contains the rules:

$$\text{parent}(x, y) \rightarrow \text{child}(y, x) \text{ and}$$
$$\text{child}(x, y) \rightarrow \text{parent}(y, x)$$

we would not be able to assume anything under most specific abduction.

3.2.5 Levesque's Knowledge-Level Account of Abduction

Levesque [9] presents an account of abduction which, in a sense, subsumes the approaches we have seen so far. Here, abduction is based upon an underlying model of belief. By varying this notion of belief, different types of abduction can be characterised. For instance, if the notion of belief used is closed under ordinary logical consequence, then the result will be the idealised logic form of abductive inference. Levesque's motivation seems to be a desire to move away from notions of abduction based solely on traditional logical implication. He cites the following example: if we know that Marc is 3 or 4 years old, then the fact that he is not yet 4 does not explain why he is 3 even though it does imply it.

We shall use a standard propositional language $\mathcal{L}$ with a special constant denoting falsity. The symbols p, q, r,... denote propositional letters of $\mathcal{L}$; α, β, γ,... denote sentences of $\mathcal{L}$; m denotes literals of $\mathcal{L}$; m represents the complement of m; and, use $\wedge\{a_i\}$ and $\vee\{a_i\}$ to, respectively, conjoin and disjoin a set of sentences. The logical language $\mathcal{L}^*$ will be used to talk about beliefs. It resembles $\mathcal{L}$ in all respects except that atomic sentences are of the form $\mathbf{B}\alpha$ (where α is a sentence of $\mathcal{L}$). Different kinds of belief will be expressed with the use of a subscript. For instance, $\mathbf{B}_\lambda\alpha$ represents a belief of type λ. We shall not talk about beliefs of beliefs.

$\mathcal{L}$ and $\mathcal{L}^*$ are interpreted in the usual manner. We shall just note that for atomic sentences of $\mathcal{L}^*$ an epistemic state e represents the sentences of $\mathcal{L}$ that are believed. The expression $e \models \mathbf{B}_\lambda\alpha$ denotes the fact that $\mathbf{B}_\lambda\alpha$ is true in epistemic state e. Also, if α is a sentence of $\mathcal{L}$ then $||\alpha||$ denotes the proposition expressed by α which we can take to be the set of assignments where α is true. We define $||\{\alpha_1, \ldots, \alpha_n\}||$ to be $\{||\alpha_1||, \ldots, ||\alpha_n||\}$.

With this language we can now define the notion of explanation for a type of belief λ with respect to an epistemic state e.

Definition 3.8 :

α expl_λ β *wrt e* iff $e \models [\mathbf{B}_\lambda(\alpha \rightarrow \beta) \wedge \neg\mathbf{B}_\lambda\neg\alpha]$.

Basically, this definition says that α is an acceptable explanation for β iff we believe that α implies β and we have no reason for believing that α is not the case. There are a few interesting points that can be made about this definition. Firstly, trivial explanations such as α expl_λ α wrt e will be permitted if $(\alpha \rightarrow \alpha)$ is believed and $\neg\alpha$ is not believed. In fact, if the only beliefs we have about β are logical truths (tautologies), then only trivial explanations will exist. Also, with some types of belief if we have $\mathbf{B}\neg\beta$ then no explanations will exist, while if we have $\mathbf{B}\beta$ then the only explanation is $\neg$.

In order to exclude explanations that contain superfluous information, we introduce the notion of a simplicity criterion.

Definition 3.9 :

The literals of α, $LITS(\alpha)$, is defined by:

$LITS(\quad) = \emptyset$;

$LITS(p) = \{p\}$;

$LITS(\neg\alpha) = \{\bar{m} \mid m \in LITS(\alpha)\}$;

$LITS(\alpha \wedge \beta) = LITS(\alpha \vee \beta) = LITS(\alpha) \cup LITS(\beta)$.

Definition 3.10 :

α is simpler than β (written $\alpha < \beta$) iff $LITS(\alpha) \subset LITS(\beta)$; also $\alpha \leq \beta$ iff $LITS(\alpha) \subseteq LITS(\beta)$.

Intuitively, this notion says that a sentence is simpler than another if it has fewer propositional letters of the same polarity.

Example 3.5 :

$(p \wedge \neg q) < (q \rightarrow (p \vee r))$, but

$(P \wedge \neg q) \nless (\neg q \rightarrow (p \vee r))$.

This notion of simplicity is not very complex and it may be that a better definition is possible but it does serve the purpose of excluding redundant hypotheses which are likely to be unrelated to what is being explained.

We are now in a position to define a minimal explanation.

Definition 3.11 :

α min–expl$_\lambda$ β wrt e iff α expl$_\lambda$ β wrt e and for no $\alpha^* < \alpha$ is it the case that α^* expl$_\lambda$ β wrt e.

There may of course be many minimal explanations, so the notion of abduction described will return all of them.

Definition 3.12 :

$EXPLAIN_\lambda[e, \beta] = ||\{\alpha \mid \alpha$ min–expl$_\lambda$ β wrt $e\}||$.

The set of explanations produced can be considered disjunctively.

Notice that this approach is independent of the representation of the epistemic state e. In his presentation, though, Levesque does not commit himself to his definitions of explanation and simplicity and mentions that other definitions may well be possible but leaves them for investigation at a later date.

3.3 Selecting the Best Hypothesis

3.3.1 Probabilistic Measures

Probabilistic measures are one approach to the problem of determining the best or most likely explanation. These methods are usually based on Bayes' rule which was introduced in Chapter 2. As noted there, due to the large number of probabilities required for a problem, it is very often the case that simplifying assumptions and heuristics are adopted to reduce the complexity of calculating the probability of an explanation. We shall not elaborate on any specific probabilistic approaches here but note that many of the simplifying assumptions and heuristics used have been criticised for being unrealistic.

3.3.2 Cost-based Measures

Cost-based measures for determining the best explanation usually work by associating a cost with making an assumption or using an axiom. Stickel [23] presents one such method. Every axiom A is assigned a cost $cost(A) > 0$ and every assumable literal L is assigned a cost $cost(L) > 0$. Recall that Stickel considers abduction to be a proof that a goal follows from substitution instances S of axioms and substitution instances H of assumable literals. The cost of a proof (explanation) can then be calculated by:

$$\sum_{A\sigma \in S} cost(A) + \sum_{L\sigma \in H} cost(L).$$

The best explanation is then considered to be that with the least cost.

Another method mentioned by Stickel [23] propagates assumption costs from consequents to antecedents. For example, if we have the implication $P_1^{w_1} \wedge P_2^{w_2} \rightarrow Q$ and if the cost of assuming Q is c, then P_1 is assumable with cost $w_1 c$ and, likewise, P_2 is assumable with cost $w_2 c$. Note that most specific abduction is preferable if $w_1 + w_2 < 1$ since it will cost less to assume P_1 and P_2 than to assume Q and that least specific abduction is preferable if $w_1 + w_2 > 1$ for similar reasons. If, say, P_1 is provable though, then assuming P_2 may be favoured depending on its weight compared to that of Q. It has been shown [3] that this method does have a probabilistic interpretation.

This method is rather straightforward though and, as with probabilistic approaches, it suffers from the disadvantage that it is difficult to attach a semantics to a cost measure. The amount of time required to compute the minimum cost explanation may also be considerable.

3.3.3 Simplicity Measures

The term "simplicity" is probably misleading as it can apply to a number of approaches to determining the best hypothesis (e.g. cost-based measures can also be considered a form of simplicity criterion) and we have already come across one notion of simplicity by Levesque [9]. Notions of simplicity are usually an evaluation criterion based on the maxim of Occam's Razor.

Ram and Leake [21] mention a few such criteria. Firstly, the explanation whose causal chain is shortest in overall length may be preferred. Alternatively, we may prefer the explanation which requires fewer assumptions. Also, if one hypothesis subsumes another then the more general one is preferred as it is likely to explain more. Charniak [2] adopts the hypothesis which maximises the expression E/A where E is the number of explained observations and A is the number of assumptions made (this would seem to be a combination of the previous two criteria).

These criteria have the advantage of being rather simple. In fact, it would seem that they are too simple as more complex criteria are being put forward (as we shall see) to account for situations where the above criteria have proved ineffective.

3.3.4 Explanatory Coherence

The term "coherence" relates to how things "stick together". Notions of *explanatory coherence* attempt to determine the best explanation by measuring how well the components of an explanation "stick together".

3.3.4.1 Ng and Mooney

Ng and Mooney [13] propose an *explanatory coherence metric* based on the proof graph for an explanation. This metric attempts to favour an explanation possessing certain properties. Explanations with more connections between any two observations and fewer disjoint partitions are especially favoured. The metric can be formally defined by the following formula

$$C = \frac{\sum_{1 \le i < j \le l} N_{i,\,j}}{N \binom{l}{2}}$$

where

l = total number of observations;

N = total number of nodes in proof graph;

$\binom{l}{2} = l(l-1)/2$; and,

$N_{i,\,j}$ = number of distinct nodes n_k in proof graph such that there is a sequence of directed edges from n_k to n_i and a sequence of directed edges from n_k to n_j, where n_i and n_j are observations.

The denominator scales the result so that it will lie within the range [0, 1]. The numerator measures the total number of nodes in an explanation connecting pairs of observations and increases with the number of nodes in an explanation that simultaneously lend support to a given connection. Ng and Mooney claim that, using a depth-first search algorithm, this value can be calculated in $O(lN + e)$ time where l and N are as above and e is the total number of directed edges in the proof graph.

The following example illustrates this coherence metric and compares it to Charniak's simplicity criterion([2] and also mentioned previously in the section on simplicity measures).

Example 3.6 :

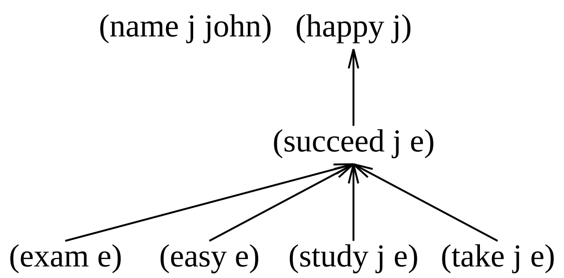
Suppose we have a knowledge base containing the following axioms:

$$(\mathit{happy}\ ?x) \leftarrow (\mathit{optimist}\ ?x)$$

$$(\mathit{happy}\ ?x) \leftarrow (\mathit{succeed}\ ?x\ ?y)$$

$$(\mathit{succeed}\ ?x\ ?y) \leftarrow (\mathit{exam}\ ?y)\ (\mathit{easy}\ ?y)\ (\mathit{study}\ ?x\ ?y)\ (\mathit{take}\ ?x\ ?y)$$

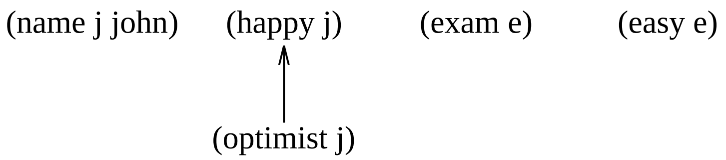
The sentence "John was happy. The exam was easy" can be represented by conjoining the literals: $(\mathit{name}\ j\ \mathit{john})$, $(\mathit{happy}\ j)$, $(\mathit{exam}\ e)$ and $(\mathit{easy}\ e)$. Using the knowledge base let us consider the following two interpretations.



$E = 1, A = 2$

Simplicity metric = $E/A = 1/2$

Coherence metric = $(1 + 1)/(7 * (4 * 3/2)) = 0.048$



$E = 1, A = 1$

Simplicity metric = $E/A = 1/1$

Coherence metric = 0

Figure 2. Explanatory coherence metric compared with a simplicity measure.

The coherence metric prefers the first interpretation which would appear to be better, while the simplicity measure prefers the second interpretation.

This approach has the advantage of being less affected by changes in the underlying representation of the knowledge base than simplicity criteria. This is because if we change the way the axioms in the knowledge base are encoded then the simplicity measure will obviously be directly affected but the connections between propositions are likely to be affected to a lesser extent. Ng and Mooney also point out that coherent explanations, being tightly connected, have a greater likelihood of assumptions being unified which results in fewer assumptions being made. Therefore, coherent explanations are likely to be simple. They also mention that coherence can be used to determine how specific an explanation should be (c.f. Stickel’s forms of abduction presented in an earlier section). In this case, a rule is used in an explanation if it leads to a increase in the value of the coherence metric.

3.3.4.2 Thagard

Thagard [25], [26] presents a much more comprehensive theory of explanatory coherence based on seven principles. This theory of explanatory coherence allows us to develop a notion of *acceptability* of a proposition. A proposition that is coherent with our beliefs should be accepted; a proposition that is incoherent with our beliefs should be rejected; and, a proposition which is neither coherent nor incoherent with our beliefs should be treated indifferently. We shall firstly present Thagard’s seven principles of explanatory coherence before briefly discussing each one in turn.

An explanatory system S consists of the propositions $P,\ Q,\ P_1,\ldots,P_n$. The notion of local coherence is a relation between two propositions. With this in mind, the seven principles of explanatory coherence are:

Principle 1. Symmetry

a. If P and Q cohere, then Q and P cohere.

b. If P and Q incohere, then Q and P incohere[6].

6. Note that if two propositions $P,\ Q$ do not cohere then they do not necessarily incohere.

Principle 2. Explanation

If $P_1 \ldots P_m$ explain Q, then:

a. For each P_i in $P_1 \ldots P_m$, P_i and Q cohere.

b. For each P_i and P_j in $P_1 \ldots P_m$, P_i and P_j cohere.

c. In (a) and (b) the degree of coherence is inversely proportional to the number of propositions $P_1 \ldots P_m$.

Principle 3. Analogy

a. If P_1 explains Q_1, P_2 explains Q_2, P_1 is analogous to P_2, and Q_1 is analogous to Q_2, then P_1 and P_2 cohere, and Q_1 and Q_2 cohere.

b. If P_1 explains Q_1, P_2 explains Q_2, Q_1 is analogous to Q_2, but P_1 is disanalogous to P_2, then P_1 and P_2 incohere.

Principle 4. Data Priority

Propositions that describe the results of observations have a degree of acceptability on their own.

Principle 5. Contradiction

If P contradicts Q, then P and Q incohere.

Principle 6. Acceptability

a. The acceptability of a proposition P in a system S depends on its coherence with the propositions in S.

b. If many results of relevant experimental observations are unexplained, then the acceptability of a proposition P that explains only a few of them is reduced.

Principle 7. System Coherence

The global explanatory coherence of a system S of propositions is a function of the pairwise local coherence of those propositions.

Principle C. Competition

> If *P* and *Q* both explain evidence *E*, and if *P* and *Q* are not explanatorily connected, then *P* and *Q* incohere. Here *P* and *Q* are explanatorily connected if any of the following conditions holds:
>
> a. *P* is part of the explanation of *Q*,
>
> b. *Q* is part of the explanation of *P*,
>
> c. *P* and *Q* are together part of the explanation of some proposition *R*,
>
> d. *P* and *Q* are both explained by some higher-level proposition *R*.

The first principle maintains that pairwise (local) coherence and incoherence are symmetric relations. Intuitively, this would appear to make sense if we recall that coherence means "stick together". Note that this is a different behaviour to that exhibited by entailment and conditional probability relations.

The second principle provides the core of the definition of explanatory coherence. Part (a) expresses the fact that a proposition that helps explain, coheres with the proposition explained. Here it is necessary that the notion of explanation be more restrictive than mere logical implication otherwise we could show that any two propositions cohere by introducing expression like $((P_1 \wedge (P_2 \wedge \neg P_2)) \rightarrow Q)$. Part (b) says that any two propositions cohere if they both form part of an explanation for another proposition. Here, too, it is necessary that the notion of explanation not coincide with logical implication otherwise any proposition that explains something would cohere with every other proposition due to the fact that if $(P_1 \rightarrow Q)$ then $(P_1 \wedge P_2 \rightarrow Q)$ and so on. Note that this is similar to notions of logical simplicity (c.f. Levesque's [9] notion of simplicity mentioned in an earlier section) that attempt to rule out explanations which make assumptions that are unrelated to what is being explained. In part (c) we see that the measure of coherence decreases as the number of hypotheses made in an explanation increases. This means that explanations requiring fewer hypotheses are favoured (c.f. simplicity measures mentioned in a previous section). Thagard actually maintains that this notion of simplicity is the most appropriate for scientific theory choice.

Principle 3 relates the notion of analogy to that of explanatory coherence. Thagard claims that analogy has been important in justifying some theories (e.g.

Darwin's theory of evolution). Part (a) expresses the fact that an explanatory analogy, not just any analogy between propositions, must exist for propositions to cohere. Part (b) tells us that two hypotheses incohere if they are dissimilar but explain analogous propositions. Thagard claims that this was one factor which assisted Einstein in proposing his theory of special relativity.

The fourth principle claims that observations should have a certain amount of acceptance on their own. This is not saying that observations should be accepted outright but rather that observations are more likely to be true and are usually believed unless there is sufficient evidence to suggest otherwise.

Principle 5 says that two propositions incohere if they contradict each other. The notion of contradiction here refers not only to propositions that logically contradict (e.g. $P \wedge \neg P$) but also to hypotheses that are incompatible and compete with each other to explain some proposition.

The sixth principle outlines a notion of acceptability of a proposition based on the notion of explanatory coherence as presented in the foregoing principles. Part (a) says that the acceptability of a proposition is dependent on the local coherence of all other propositions in the explanatory system. Part (b) claims that the acceptability of a hypothesis should be reduced if there is no hypothesis that can account for much of the relevant evidence (i.e. the best explanation is still not very good). Relevant evidence refers to that which is explained by a hypothesis or one of its competitors or to evidence which is similar to this type of evidence.

The seventh principle suggests that we can have a measure of the global coherence of an explanatory system and that it will rely on the local coherences. The principles above tell us that an explanatory system will have a higher global coherence if:

i. it has more data;

ii. it has more internal explanatory links between propositions that cohere because of explanations and analogies; and,

iii. it succeeds in separating coherent subsystems of propositions from conflicting subsystems.

Principle C was introduced by Thagard [26] to improve the theory's handling of some cases. It says that, unless we have evidence to the contrary, we should assume that two hypotheses that explain the same evidence compete with each other. This suggests that two propositions may incohere even if there is no special relationship between them[7].

Thagard also presents a connectionist algorithm that implements these principles. Nodes (units) in the connectionist network represent propositions in the explanatory system. Each unit has an associated degree of activation which represents the propositions acceptability. If two propositions cohere, an excitatory link is placed between their respective units (the links can be considered bidirectional). On the other hand, if two propositions incohere, an inhibitory link is placed between their respective units. Each link has an associated weight which represents the degree of coherence or incoherence of the propositions (e.g. excitatory links will have positive weights while inhibitory links have negative weights). A special evidence unit is used to represent data and is given an initial value to signify its initial acceptability.

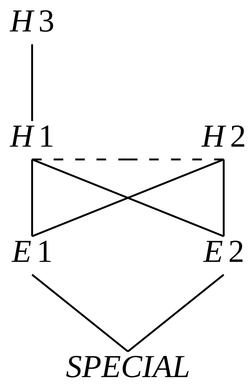
Example 3.7 :

If we have the following input

$(EXPLAIN\ (H1)\ E1)$
$(EXPLAIN\ (H1)\ E2)$
$(EXPLAIN\ (H2)\ E1)$
$(EXPLAIN\ (H2)\ E2)$
$(EXPLAIN\ (H3)\ H1)$
$(EXPLAIN\ (H1)\ E1)$
$(CONTRADICT\ H1\ H2)$
$(DATA\ (E1\ E2))$

then the following network will be produced (note that weights on links and degrees of activation have been omitted).

7. Note the change from Principle 5.



Once this network is constructed, it can be "run". Activation will spread from the data units to the other units in the network. Propositions that are connected by excitatory links will "help" each other while those connected by inhibitory links will tend to "hinder" each other (this of course would change for units with negative activations). Once the activation values reach an asymptote then the network can be "stopped". Propositions with positive activation can be considered accepted while those with negative activation are considered rejected. If the activation is near zero then the proposition is neither accepted nor rejected. Thagard claims that this network actually assigns higher activations to hypotheses that explain more, provide simpler explanations and are analogous to other hypotheses.

This theory of explanatory coherence has been used to model competing scientific theories (e.g. phlogiston theory vs. Lavoisier's theory of combustion and collision theory vs. climatic theory of extinction of dinosaurs) and legal arguments among others and the results have been promising. The computational model used seems to provide a natural interpretation for some of the notions proposed by the principles of explanatory coherence. It also possesses an interesting property in that new information can be added to the network and it will resettle into the expected result. Another advantage of this approach is that explanations need not account for all observations. It is possible that hypotheses supporting a certain theory will be acceptable even if they do not explain all the evidence. As mentioned previously, this is a common occurrence, even in scientific fields.

One problem with this approach, though, is that it provides no real definition of notions such as explanation and analogy. Thagard says that he will not "offer a theory of explanation"([25] p436) but claims that it is compatible with other accounts of explanation. In particular, he mentions that explanation is sometimes considered deductive, yet as we have seen, Principle 2 rules out this possibility. Clearly, a notion of explanation is needed (as are other notions, such as analogy) since the principles presented are making implicit assumptions regarding its nature.

3.3.5 Utility-based Measures

Many of the measures for determining the best hypothesis that we have discussed so far have been directly related to the structural (syntactic) properties of an explanation. Ram and Leake [21] introduce the notion of utility-based criteria which select hypotheses based on a system's intended use for an explanation. Their main use of these criteria are in explanations of anomolies (i.e. situations where our expectations were not realised).

Ram and Leake claim that an explanation of an anomoly must answer two questions:

i. Why did things occur as they did in the world? This question focuses on understanding and learning about the causal structure of the domain.

ii. Why did I fail to predict this correctly? This question focuses on understanding and improving the organisation of the reasoner's own model of the domain.

Utility-based criteria are methods of evaluation related to these two criteria. Utility-based criteria refer to any methods of evaluating explanations based on goals arising from these two questions. In order to clarify the form these criteria may take, we shall briefly discuss these questions.

The first of theses questions relates to the causality of the domain and is thus called a domain *domain explanation*. It leads to knowledge acquisition goals which aim to collect more information about the domain. These goals are usually guided by the explainer's tasks. Some of the tasks that can give rise to knowledge acquisition goals are:

- Choosing a response to an unexpected event — learn causes that allow discrimination between possible plans by predicting events or identifying current circumstances

- Repairing an undesirable state — learn repairable causes of that state
- Causing recurrence — learn achievable causes
- Preventing recurrence — learn blockable causes
- Assigning credit or blame — learn particular actors' influence on an outcome
- Replicating another actor's success — learn motivations of the observed actor's unusual planning decisions.

The second question involves the reasoning processes of the system and is called *introspective* or *meta-explanation*. It results in knowledge organisation goals which aim to improve the organisation of knowledge in memory. Some of the factors that can characterise this type of goal are:

- Missing knowledge — learn new knowledge to fill gap in domain model
- Unconnected knowledge — learn new connection or new index
- Implicit assumption — learn heuristics for when to check assumption explicitly
- Calculated simplification — learn heuristics for when to check assumption in detail
- Explicit assumption — learn new knowledge to correct the assumption
- Conjunctive assumptions — learn new interactions.

4 Conclusion

Abductive inference is beginning to receive increased attention as an important form of reasoning, particularly within the field of Artificial Intelligence. It has been applied to problems in a number of areas, including diagnostic tasks([20], [12], [14]), database updates([7]), scientific discovery and learning([11]) and natural language interpretation([2], [3], [13], [23]). Unlike traditional approaches, such as rule-based systems (e.g. MYCIN and PROSPECTOR — refer to [2] for more information) which are based on deduction, abductive approaches have the advantage of being a more natural way of solving these problems.

In performing abductive inference, there are two important aspects to consider: constructing explanations; and, determining which of these constitutes the best explanation. In this report we have considered various attempts at providing a formal framework to express these notions. These approaches can be divided into two broad categories: set-cover based approaches and logic based approaches. Levesque [9] notes that set-cover based approaches suffer from difficulty in being able to express how a minor alteration to the background knowledge can contribute to changing what constitutes an explanation. Logic based approaches, on the other hand, tend to confine reasoning to global properties of logic like consistency and implication[8]. Also, logic based approaches are usually more general than set-based approaches.

There is little doubt that abduction is an important form of reasoning alongside deduction and induction. Some philosophers though, argue against Peirce's claim that there is a logic of discovery. Peirce himself was unable to provide an adequate justification for his theory of abductive inference (refer to [6] for more details). But,

8. Levesque's knowledge level approach [9] goes some way to solving this problem.

based on some of the results achieved so far, it would appear that his faith in abduction was justified.

5 Bibliography

1. Allemang, D., Tanner, M.C., Bylander, T., and Josephson, J. Computational Complexity of Hypothesis Assembly. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, Milan, 1987, pp. 1112-1117.

2. Charniak, E. Motivation Analysis, Abductive Unification, and Nonmonotonic Equality. *Artificial Intelligence 30* (1987), 275-295.

3. Charniak, E. and Shimony, S.E. Probablistic Semantics for Cost Based Abduction. In *Proceedings of the National Conference in Artificial Intelligence*, 1990, pp. 106-111.

4. Cox, P.T. and Pietrzykowski, T. Causes for Events: Their Computation and Applications. In *Proceedings of the Eighth Conference on Automated Deduction*, 1986, pp. 608-621.

5. Eshghi, K. and Kowalski, R.A. Abduction Compared with Negation by Failure. In *Proceedings of the Sixth International Conference on Logic Programming*, 1989, pp. 234-254.

6. Fann, K.T. *Peirce's Theory of Abduction*. Martinus Nijhoff, The Hague, Holland, 1970.

7. Kakas, A.C. and Mancarella, P. Database Updates Through Abduction. In *Proceedings of the 16th VLDB Conference*, Brisbane, Australia 1990, pp. 650-661.

8. Konolige, K. A General Theory of Abduction. In *Working Notes of the 1990 Spring Symposium on Automated Abduction*, P. O'Rorke, Ed. Technical Report 90-32, University of California, Irvine, September 27, 1990, pp. 62-66.

9. Levesque, H.J. A Knowledge-Level Account of Abduction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 1989, pp. 1061-1067.

10. Lewis, C. and Mack, R. The Role of Abduction in Learning to Use a Computer System. Report RC 9433, IBM Research, Yorktown Heights, 1982.

11. Morris, S. and O'Rorke, P. An Approach to Theory Revision Using Abduction. In *Working Notes of the 1990 Spring Symposium on Automated Abduction,* P. O'Rorke, Ed. Technical Report 90-32, University of California, Irvine, September 27, 1990, pp. 33-37.

12. Nau, D.S. and Reggia, J.A. Relationships Between Deductive and Abductive Inference in Knowledge-Based Diagnostic Problem Solving. In *Proceedings of the First International Workshop on Expert Database Systems*, Kerschberg, Ed., 1984, pp. 500-509.

13. Ng, H.T. and Mooney, R.J. On the Role of Coherence in Abductive Explanation. In *Proceedings of the National Conference in Artificial Intelligence*, 1990, pp. 337-342.

14. Peng, Y. and Reggia, J.A. *Abductive Inference Models for Diagnostic Problem-Solving*. Springer-Verlag, 1990.

15. Poole, D. A Logical Framework for Default Reasoning. *Artificial Intelligence 36* (1988), 27-47.

16. Poole, D. Explanation and Prediction: an Architecture for Default and Abductive Reasoning. *Computational Intelligence 5* (1989), 97-110.

17. Poole, D. On the Comparison of Theories: Preferring the Most Specific Explanation. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1985, pp. 144-147.

18. Poole, D. Representing Diagnostic Knowledge for Probabilistic Horn Abduction. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, Sydney, 1991, pp. 1129-1135.

19. Poole, D., Goebel, R., and Aleliunas, R. Theorist: A Logical Reasoning System for Defaults and Diagnosis. In *The Knowledge Frontier: Essays in the Representation of Knowledge*, N. Cercone and G. Mc Calla, Eds. Springer Verlag, 1987, pp. 331-352.

20. Pople, H.E. On the Mechanisation of Abductive Logic. In *Proceedings of the Third International Joint Conference on Artificial Intelligence*, Vol. 2, 1973, pp. 147-152.

21. Ram, A. and Leake, D. Evaluation of Explanatory Hypotheses. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, Chicago, IL, August 1991, pp. 867-871.

22. Sowa, J.F. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, 1984.

23. Stickel, M.E. A Prolog-like Inference System for Computing Minimal-Cost Abductive Explanations in Natural-Language Interpretation. *Annals of Mathematics and Artificial Intelligence 4* (1991), 89-106.

24. Thagard, P. Review of *Abductive Inference Models for Diagnostic Problem-Solving* by Y. Peng and J.A. Reggia Springer-Verlag. *SIGART Bulletin 2*, 1 (January, 1991), 72-75.

25. Thagard, P. Explanatory Coherence. *Behavioral and Brain Sciences 12* (1989), 435-502.

26. Thagard, P. The Dinosaur Debate: Explanatory Coherence and the Problem of Competing Hypotheses. In *Philosophy and AI: Essays at the Interface*, J. Pollock and R. Cummins, Eds. MIT Press/Bradford Books, To Appear.

Approaches to

Abductive Inference

(The Never Ending Search for the Best Explanation)

by

Maurice Pagnucco

Basser Department
of Computer Science†

University of Sydney

March, 1992

†This work was performed within the Knowledge Systems Group.

ABSTRACT

Abductive reasoning is a fundamental form of logical inference which aims to derive plausible explanations for a body of data. It works by generating hypotheses which would explain a group of facts (usually observations). The term "abduction" was first introduced by the American philosopher Charles Sanders Peirce who was also the first person to distinguish it as a fundamental form of reasoning alongside deduction and induction. Peirce considered abduction to be a method of proposing hypotheses in scientific discovery and maintained that it represented the first stage of scientific inquiry.

Abduction is beginning to receive increased attention as an important form of reasoning, particularly within the field of Artificial Intelligence. It has been applied to a variety of problems, including diagnostic tasks, database updates, scientific discovery and learning and natural language interpretation. Unlike traditional approaches to some of theses problems, such as rule-based systems which, typically, are based on deduction, abduction has the advantage of providing a more natural problem solving approach.

Two major factors must be considered when investigating abduction: constructing an explanation; and, selecting the "best" or most plausible explanation from among these. In this report we investigate a number of attempts at providing a formal framework within which to express these notions. Currently, these approaches can be divided into two broad categories: set-cover based and logic based approaches. Here, we shall consider methods from both of these areas along with a brief discussion of Peirce's thoughts on abduction.

Keywords:

Abduction, Explanation, Reasoning, The albatross flies in the night sky, Set-based approaches, The pig ruts in the farmyard, Logic based approaches.

ACKNOWLEDGEMENTS

<Add Name Here>

CONTENTS

LIST OF FIGURES