EXPLANATION AND PREDICTION: AN ARCHITECTURE FOR DEFAULT AND ABDUCTIVE REASONING

by

David Poole

Technical Report 89-4

Explanation and Prediction: An Architecture for Default and Abductive Reasoning

David Poole
Department of Computer Science,
The University of British Columbia,
Vancouver, B.C., Canada V6T 1W5
(604) 228 6254
poole@cs.ubc.ca

March 20, 1989

Abstract

Although there are many arguments that logic is an appropriate tool for artificial intelligence, there has been a perceived problem with the monotonicity of classical logic. This paper elaborates on the idea that reasoning should be viewed as theory formation where logic tells us the consequences of our assumptions. The two activities of predicting what is expected to be true and explaining observations are considered in a simple theory formation framework. Properties of each activity are discussed, along with a number of proposals as to what should be predicted or accepted as reasonable explanations. An architecture is proposed to combine explanation and prediction into one coherent framework. Algorithms used to implement the system as well as examples from a running implementation are given.

Key words: defaults, conjectures, explanation, prediction, abduction, dialectics, logic, nonmonotonicity, theory formation

1 Introduction

One way to do research in Artificial Intelligence is to argue that we need a certain number of tools and to augment these only when they have proven inadequate for some task. In this way we can argue we need at least the first order predicate calculus to reason about individuals and relations among individuals (given that we want to indirectly describe individuals, as well as describe the conjunction, disjunction and negation of relations) [Hayes77,Moore82,Genesereth87].

Non-monotonicity has often been cited as a problem with using logic as a basis for commonsense reasoning. In [PGA87] it was argued that instead of deduction from our knowledge, reasoning should be viewed as a process of theory formation. In [Poole88a] it was shown how default reasoning can be viewed in this way by treating defaults as possible hypotheses that can be used in an explanation.

It has also been recognised (e.g., [Charniak85,PGA87,Cox87,Reggia83]) that abduction is an appropriate way to view diagnostic and recognition tasks. In diagnosis, for example, the diseases and malfunctions are the possible hypotheses that can be used to explain some observations.

We can argue we want to use logic and do hypothetical reasoning. This research considers the simplest form of hypothetical reasoning, namely the case where the user provides a set of possible hypotheses they are prepared to accept as part of a theory. This is the framework of the Theorist system [Poole88a,PGA87]. The distinctions outlined in this paper were found from experience by using the system, explaining to others how to use the system and in building applications [Poole87b].

1.1 Theorist Framework

We assume we are given a standard first order language over a countable alphabet [Enderton72]. By a formula we mean a well formed formula in this language. By an instance of a formula we mean a substitution of terms in this language for free variables in the formula. In this paper the Prolog convention of variables starting with an upper case letter is used.

The framework [Poole88a] is defined in terms of two sets of formulae:

$\mathcal{A}$ is a set of closed formulae which we are taking as given, and

H is a set of (possibly open) formulae which we take as the "possible hypotheses".

Definition 1.1 *A* **scenario** *of* (A, H) *is a set* D *of ground instances of elements of* H *such that* $D \cup A$ *is consistent.*

A scenario is a set of hypotheses that could be true based on what we are given.

Definition 1.2 *If* g *is a closed formula, an* **explanation** *of* g *from* (A, H) *is a scenario of* (A, H) *which (together with* A*) implies* g*.*

Thus D is an explanation of g from (A, H) if D is a set of ground instances of elements of H such that $D \cup A$ is consistent and $D \cup A \models g$.

Definition 1.3 *An* **extension** *of* (A, H) *is the set of logical consequences of* A *together with a maximal (with respect to set inclusion) scenario of* (A, H)*.*

The following theorem was proved in [Poole88a] and follows from the compactness theorem of the first order predicate calculus [Enderton72].

Theorem 1.4 *There is an explanation of* g *from* A, H *iff* g *is in some extension of* A, H*.*

In [Poole88a] it was shown that $\delta \in H$ corresponds exactly to the normal default $: \delta / \delta$ of [Reiter80]. It was also argued that the extra power of Reiter's defaults was not needed. Both [Reiter80] and [Poole88a] showed how their systems can be used for default reasoning, but not what such reasoning was for.

1.2 Explanation and Prediction

There are two activities we will consider, namely explaining observations and predicting what is expected to be true. These are both considered to be instances of the Theorist framework.

I make the assumption that we do not need more than the Theorist framework. This may turn out to be incorrect, but if it is, we will have found a good reason to add extra features to our system. I make no a priori assumption that the same hypotheses should be used both for explanation and prediction; in fact there are good reasons for not making them the same. If we later find out they coincide, again we will have learnt something.

As such, the following sets of formulae are provided by the user:[1]

F is the set of *facts*, which are taken as being true of the domain;

Δ is the set of *defaults*, possible hypotheses which can be used in prediction;

Π is the set of *conjectures*, possibly hypotheses which can be used in explaining observations;

O is the set of *observations* that have been made about the actual world.

2 Prediction

A problem many people in AI have been working on is the problem of predicting what one expects to be true in some (real or imaginary) world based on the information one has about that world.

The most conservative form of prediction is logical consequence from our knowledge. If axioms A are true in some world, any logical consequence of A must also be true in that world. This is the essence of classical logic.

Many people have argued that such a notion is too weak for common sense prediction; sometimes we want to make assumptions as to what we expect to be true. This is the basis of much work on nonmonotonic reasoning [Bobrow80].

We consider defaults as assumptions one is prepared to make about the world, unless they can be shown to be wrong. In the Theorist framework, defaults are possible hypotheses used for prediction [Poole88a].

What should be predicted based on such hypothetical reasoning seems to be uncontroversial if there are no conflicting defaults (i.e., there is only one extension). In this section, we discuss what should be predicted when there are conflicting defaults.

[1] As far as the preceding semantics are given, the possible hypotheses, H, will in some cases be Δ and in some cases $\Pi \cup \Delta$; the given A will sometimes be F and sometimes F together with an explanation of the observations.

We assume there is no other information on which to base our decision (e.g., specificity [Poole85], probability [Neufeld87], temporal considerations [Goebel87]).

Example 2.1 Consider the following example[2]

$$\begin{aligned} H \ = \{ \ & \mathit{republican}(X) \Rightarrow \mathit{hawk}(X), \\ & \mathit{quaker}(X) \Rightarrow \mathit{dove}(X), \\ & \mathit{hawk}(X) \Rightarrow \mathit{support\text{-}star\text{-}wars}(X), \\ & \mathit{hawk}(X) \Rightarrow \mathit{politically\text{-}motivated}(X), \\ & \mathit{dove}(X) \Rightarrow \mathit{politically\text{-}motivated}(X) \\ & \mathit{quaker}(X) \Rightarrow \mathit{religious}(X)\} \\ F \ = \{ \ & \forall X \ \neg(\mathit{dove}(X) \wedge \mathit{hawk}(X)), \\ & \mathit{quaker}(\mathit{dick}), \\ & \mathit{republican}(\mathit{dick}) \ \} \end{aligned}$$

Based on the above facts and defaults, there are questions as to which of the following should be predicted:

$$\begin{aligned} & \mathit{dove}(\mathit{dick}) \\ & \mathit{hawk}(\mathit{dick}) \\ & \mathit{dove}(\mathit{dick}) \vee \mathit{hawk}(\mathit{dick}) \\ & \mathit{dove}(\mathit{dick}) \wedge \mathit{hawk}(\mathit{dick}) \\ & \mathit{support\text{-}star\text{-}wars}(\mathit{dick}) \\ & \mathit{politically\text{-}motivated}(\mathit{dick}) \\ & \mathit{religious}(\mathit{dick}) \end{aligned}$$

The rest of this section discusses four proposals of what should be predicted. There are based on the answers to the following question:

If we have an explanation for p and and an explanation for q, but we know both cannot be true (i.e., $F \models \neg(p \wedge q)$), what should we predict?

1. Either p or q but not both.

[2] This example is based on an example by Matt Ginsberg, which is based on an example of Ray Reiter.

2. Neither p nor q.
3. $p \vee q$.
4. Nothing; we have detected an inconsistency in our knowledge base.

The following sections consider the consequences of each choice.

2.1 Predict if explainable

The first definition of prediction where we predict p or predict q corresponds to predicting whatever is explainable (predicting what is in some extension[3]).

In example 2.1, we would predict either

$$\textit{politically-motivated}(dick) \wedge hawk(dick) \wedge \textit{supports-star-wars}(dick)$$
$$\text{or}$$
$$\textit{politically-motivated}(dick) \wedge dove(dick)$$

but not both. This can be claimed to be reasonable because we were told we could assume Dick is a hawk given no evidence to the contrary (the only evidence to the contrary being an internal inconsistency), and so can conclude he is politically motivated and supports star wars. We can also assume he is a dove and so is politically motivated. We just cannot assume he is both a dove and hawk, as this is inconsistent.

This has the peculiar property that we both predict $hawk(dick)$ and predict $\neg hawk(dick)$ (although in different extensions). The following shows this turns out to be general.

Theorem 2.2 *There are multiple extensions if and only if there is some α such that α is explainable and $\neg\alpha$ is explainable.*

> **Proof:** Suppose there are two extensions, E_1 and E_2. Different extensions are mutually inconsistent, so $F \cup E_1 \cup E_2$ is inconsistent. By the compactness of the first order predicate

[3] [Reiter80] uses membership in one extension, but does not claim that he is formalising prediction, but rather "an acceptable set of beliefs that one may hold about an incompletely specified world" [Reiter80, p. 88].

calculus, there are finite subsets D_1 and D_2 of E_1 and E_2 respectively such that $F \cup D_1 \cup D_2$ is inconsistent. D_1 is such an α as D_1 is in extension E_1 and $\neg D_1$ is in E_2 (as $F \cup D_2 \models \neg D_1$).

Conversely, suppose α is explained by D_1 and $\neg\alpha$ is explained by D_2. Extend D_1 to extension E_1 and D_2 to E_2. E_1 and E_2 are mutually inconsistent, and so are different. Thus there are multiple extensions. □

As it seems wrong to both *predict* α and *predict* $\neg\alpha$, membership in an extension seems like a strange notion of prediction. It corresponds more to "may be true" than to prediction.

2.2 Incontestable Scenarios

When both p and q can be explained, but are mutually inconsistent, it seems reasonable to predict neither; we were told we could assume p given no evidence to the contrary, but q is evidence to the contrary, so we should not assume p.

This notion of prediction, corresponds to predicting what can be explained using an "incontestable scenario". This is a very sceptical form of prediction where we predict some goal only if we have an argument why the goal should be true (i.e., the goal in explainable) and we cannot find an argument why the argument for the goal should not be true.

In example 2.1, of the conclusions suggested only $religious(dick)$ is predicted. We can't assume he is hawk since, as far as we know, he could be a dove, and we can't assume he is a dove, as he may as well be hawk (and he can't be both), so nothing that depends on these is predicted.

Definition 2.3 Scenario D of (F, Δ) is an **incontestable scenario** if $\neg D$ is not explainable from (F, Δ).

The following lemma shows that being in an incontestable scenario is a local property of instances of defaults and does not depend on other defaults in an explanation.

Lemma 2.4 *Scenario D of (F, Δ) is an incontestable scenario iff for all $d \in D$, $\neg d$ is not explainable from (F, Δ).*

Proof: Scenario S explains $\neg D$ iff there is some minimal subset D' of D such that $F \cup S \models \neg D'$.

The lemma follows from noticing that if $D' = \{d_1, ..., d_n\}$,

$$F \cup S \models \neg(d_1 \wedge ... \wedge d_n)$$

iff

$$F \cup S \cup \{d_1, ..., d_{n-1}\} \models \neg d_n$$

and the left hand side of each formula is consistent (by the minimality of D'). □

Thus being part of an incontestable scenario is a local property of instances of defaults and so there is a unique incontestable extension, defined as:

Corollary 2.5 g is incontestably explainable from (F, Δ) iff g logically follows from $F \cup D$ where

$$D = \{d : d \text{ is a ground instance of an element of } \Delta \text{ and } \neg d \text{ is not explainable from } (F, \Delta)\}$$

For the ground case, if we can explain the negation of a default, it can be removed at compile time. A new knowledge base can be built by removing any default from Δ for which we can explain its negation (from F and the initial Δ), and then computing logical consequences of the facts and the remaining defaults (i.e., those for which we cannot explain their negations).

For the non-ground case, however, this does not work as we cannot remove a default just because the negation of some instance of it is explainable. In this case the set D may be infinite, however we can still check explainability dynamically.

2.3 Membership in all Extensions

The third response to the question posed in section 2 was to predict the disjunction $p \vee q$.

We do not predict something if we can just explain it, as we may be able to explain it and its negation. It seems wrong to both predict some proposition and also predict its negation. It is also not adequate to predict some proposition because we can explain it and cannot explain its negation. Consider an example where we can explain a and can also explain $\neg a$. We do not want to predict a is true. Suppose the only rule about g is $a \Rightarrow g$; if we can't predict a, we do not want to predict g, even though there is no way to explain $\neg g$. Such considerations lead to the idea of predicting what is in every extension (or, equivalently what logically follows from the disjunction of the maximal scenarios).

In this section we discuss different properties of such prediction; in section 5.2 we show how it can be implemented.

In example 2.1, we predict

$$\begin{gathered} religious(dick) \wedge politically\text{-}motivated(dick) \wedge \\ ((hawk(dick) \wedge supports\text{-}star\text{-}wars(dick)) \vee dove(dick)) \end{gathered}$$

This is the formula which is in all extensions (together with the facts, it is equivalent to the disjunction of the extensions). Whichever extension is true in a world, this formula will be true in that world.

The following theorem gives a characterisation of membership in all extensions:

Theorem 2.6 *The following are equivalent:*

1. *g is in every extension of (A, H).*
2. *for all scenarios S of (A, H), there is an explanation of g from $(A \cup S, H)$.*
3. *there does not exist a scenario S of (A, H) such that there is no explanation of g from $(A \cup S, H)$.*
4. *there is a set $\mathcal{E}$ of (finite) explanations of g from (A, H) such that there is no scenario S of (A, H) inconsistent with every element of $\mathcal{E}$.*
5. *there is an explanation D of g, and if there exists $d \in D$ such that $\neg d$ is explainable by E, then g is in every extension of $(A \cup E, H)$.*

Proof: $2 \Rightarrow 1$. If g is explainable from all scenarios, it is explainable from all maximal scenarios, that is it is in every extension.

$2 \Leftrightarrow 3$. These are rewritings of the same statement.

$4 \Rightarrow 2$. Suppose 4 holds, and there is a scenario S from which g is not explainable. Each $E \in \mathcal{E}$ is inconsistent with S (otherwise $E \cup S$ is an explanation of g from $(A \cup S, H)$).

$1 \Rightarrow 4$. Suppose 1 holds. The set of all maximal scenarios has the property given in 4 (except the finite membership). By the compactness theorem of the first order predicate calculus [Enderton72] there is a set $\mathcal{E}$ composed of finite subsets of the maximal scenarios which imply g. If some S were inconsistent with all elements of $\mathcal{E}$ it would be inconsistent with the maximal scenarios, and we know such an S cannot exist. So $\mathcal{E}$ is a set which satisfies 4.

$4 \Rightarrow 5$. Suppose 4 holds, the set $\mathcal{E}$ is countable (as it is a subset of the set of finite strings in a language with countable generators). Let D be the minimum element of $\mathcal{E}$ according to some ordering. We know $A \wedge D \models g$. As g is in every extension of A, H it is in every extension of $A \wedge E, H$.

$5 \Rightarrow 2$. Suppose 5 holds and there is some scenario S such that g is not explainable from S. D is inconsistent with S (otherwise $S \cup D$ is an scenario of S, H which explains g), so there is some $d \in D$ which follows from consistent $S' = S \cup D'$ where $D' \subseteq D$ and so by 5, g is in every extension of S', and so is in one extension of S', a contradiction to g not being explainable from S. □

This theorem shows that membership in all extensions is also a sceptical theory of prediction.

When predicting what is in all extensions, we can think of starting with all explainable propositions. We eliminate a proposition if its negation can be explained, or if its derivation rests on removed propositions. Suppose α is explainable; if $\neg\alpha$ is explainable (by scenario S), α is not in every extension. If β was derived from α, to be in all extensions β must be explainable from S .

Theorem 2.6 tells us that if g is not in every extension of A, Δ, there is some scenario S of A, Δ, such that g is not explainable from S, Δ. Based on defaults being normality conditions (i.e., conditions that we expect to be true given no evidence to the contrary) we cannot rule out S, and so we should not predict g.

The difference between predicting what is in all extensions and predicting what is incontestably explainable, is that the latter requires one explanation of the goal which is consistent with all scenarios, whereas the former allows a set of explanations of the goal which must be consistent with all scenarios.

Example 2.7 Suppose we are using the default reasoning system for recognition. Suppose also that we can explain Polly being an emu and also explain Polly being an ostrich. It cannot be both an emu and an ostrich.

$$\begin{aligned}
H &= \{ \; feathered(X) \wedge big(X) \wedge runs(X) \Rightarrow emu(X), \\
&\qquad feathered(X) \wedge big(X) \wedge runs(X) \Rightarrow ostrich(X)\} \\
F &= \{ \; \forall X \, \neg(emu(X) \wedge ostrich(X)) \\
&\qquad \forall X \; emu(X) \Rightarrow bird(X), \\
&\qquad \forall X \; ostrich(X) \Rightarrow bird(X), \\
&\qquad feathered(polly), \\
&\qquad big(polly), \\
&\qquad runs(polly)\}
\end{aligned}$$

Predicting what is incontestably explainable would not allow us to conclude anything about the identity of Polly. Neither default is usable; they effectively neutralise each other. It seems more reasonable to conclude that Polly is either an emu or an ostrich, in either case concluding Polly is a bird. This latter result is produced by membership in every extension.

If every extension contains one element of a set $\{a_i\}$ then the disjunct of the a_i is in every extension. Although scenarios are conjunctions of formulae, what is predicted is the disjunction of each extension.

In section 5.2 we show how theorem 2.6 leads to a dialectical view of prediction which can be exploited to implement membership in every extension.

2.4 Breaking Conventions

If we equate defaults with conventions, as exemplified in Autoepistemic logic [Moore85], it is reasonable that multiple extensions indicate a bug in the knowledge base [Poole89]. The "convention" view of a default says that if there is an exception to a default it must be explicitly listed. If there are multiple extensions, we should debug the knowledge base rather than solve the multiple extension problem.

If we can explain p and explain q, where p and q are mutually inconsistent, the knowledge base must have an error. One of p and q must be false in the world being axiomatised so the exception should be explicitly given in the database.

In example 2.1, the system would say that there is a bug in the database. Under the convention reading, the first default says "Unless told explicitly otherwise, if some individual is a republican they are a hawk". We know one of the first two defaults are false, so we know the user has mislead the system. The user must cancel one of the defaults [Poole88a], to say that we cannot assume Dick is a dove or we cannot assume Dick is a hawk.

Section 6.3 shows how multiple extensions can be automatically detected. This is probably useful whether or not the strict convention view of defaults is taken.

2.5 Prediction Summary

In summary, without any preference criteria for scenarios (for example [Poole85]), there is a sequence of less sceptical prediction mechanisms based on default reasoning:

1. predict only the logical consequence of the facts

2. predict what is incontestably explainable

3. predict what is in every extension

4. predict what is in any extension

5. predict what is not inconsistent with the facts

It does not seem reasonable to be less sceptical than (5) or (unless solving problems of logical omniscience) be more sceptical than (1). As discussed earlier, based on using defaults, and no preference for scenarios, it seems as though (3) is the most reasonable definition of prediction; this definition will be used for the rest of this paper.

3 Explaining Observations

When explaining actual observations, we want to build an explanation of *why* those observations could have occurred.

Dating back to C. S. Peirce's use of abduction, many people have considered the problem of finding explanations of observations. In AI there have been many abductive systems (e.g., [Reggia83,Popl73,Cebulka88,Josephson87]), but those that have been based on logic have either been based on the principle of hypothesising whatever cannot be proven (e.g., [Cox87,Popl73]) or use non-classical logics as the basis for abductive reasoning [Console89].

This section has two main aims:

1. to show that the Theorist conception of logical arguments from a predefined set of possible hypotheses is a simple, powerful and useful way to view explanation;

2. to show how explanation and prediction can be combined into one coherent framework.

The basic idea is that given some observations of the world, the system builds a theory of the world which would explain those observations. In the Theorist framework, the user provides a set of building blocks (the "possible hypotheses") from which the theory can be constructed. In diagnostic tasks the building blocks may be assumptions of normality and abnormality. In recognition tasks the building blocks are models of objects that could appear in the domain.

As prediction and explanation are different activities, I am proposing a separate set of possible hypotheses that can be used for explaining observations. These will be known as "conjectures". These are hypotheses available to explain observations, but cannot be assumed given no evidence (e.g., that some component is malfunctioning, that there is a tiger under the table).

A proposition being a default means it can be hypothesised to predict what is expected to be true. It seems reasonable to also be able to use defaults to explain observations; one explanation of an observation may be that everything is acting normally. Thus, I would expect the set of defaults to be used for explaining observations as well as for prediction, but the conjectures can only be used for explaining observations[4]. If defaults have the reading "typically", conjectures should have the reading "possibly".

Note that conjectures are different to negations of defaults from which we predict our explanations. We are not assuming that a person does not have a disease, we are just not assuming that a person has the disease. For example, if we have a set of disjoint and covering descriptors of the weather outside, we don't want to assume that the weather is not like each of them (which would be inconsistent), nor do we want to assume what the weather is like, we just want to be able to describe the weather once we encounter it. The differences between these two approaches is discussed in [Poole88c].

If we are given facts F, conjectures Π and defaults Δ, and O is observed, we want to explain O from $F, \Pi \cup \Delta$. That is, we want sets P and D, instances of elements of Π and Δ respectively, such that

$F \cup P \cup D \models O$ and
$F \cup P \cup D$ is consistent

$P \cup D$ is an explanation of O.

3.1 Existential Explanations

Consider the following example (adapted from [Kautz87]).

Example 3.1 Suppose we want to hypothesise goals for an agent, and one of the possible goals an agent can have is to go hunting with a gun in a forest. If they go hunting, they get the gun and go to the forest. This can be represented as

$$\Pi = \{hunt(W, P)\}$$

[4] The use of defaults for explaining observations is not central to the thesis of this paper; I cannot think of a case where one would not want to use them for explaining observations.

$$F = \{\forall W \forall P \; hunt(W,P) \Rightarrow get(W) \wedge goto(P)\}$$

Suppose we observe them getting gun g; there are infinitely many explanations of the form

$$\{hunt(g,\alpha)\}$$

for each ground term of the language in the place of α. This is not unreasonable, in that we want to hypothesise they are going hunting somewhere. The set of the explanations is the set of things that could be true to make the observations true. There is a difference, particularly when comparing explanations, between the infinite set of explanations represented by the schema

$$\{hunt(g,\alpha)\}$$

and the formula

$$\{\exists X \; hunt(g,X)\}$$

It seems as though the formula better represents the explanation of the observations. This is especially important when there are exceptions, for example when we know one cannot go hunting in a city park, and have the following also as facts:

$$\forall P \forall W \;\; city\_park(P) \Rightarrow \neg hunt(W,P)$$
$$city\_park(stanley\_park)$$

In the schema representation we have to list such exceptions; for the existentially quantified scenario, we do not need to consider such exceptions until we want to hypothesise a particular instance of the quantified variable.

The definition of an explanation will be extended to allow existentially quantified variables in an explanation[5]. Formally an instance of a hypothesis can be obtained by substituting any term for a variable, free variables being implicitly existentially quantified.

[5] This avoids the difficult problems that arise when we allow universally quantified variables as well as existentially quantified variables in explanations [Poole87a].

3.2 Comparators for Explaining Observations

As noticed by William of Ockham at the start of the fourteenth century, not all explanations are born equal ("What can be done with fewer [assumptions] is done in vain with more" [Edwards67, Vol. 8, p. 307]).

In this paper three different comparators for explanation, each of which could be argued for in terms of simplicity, are discussed:

1. preference for the *minimal explanation*; we prefer the explanations that makes the fewest (in terms of set inclusion) assumptions[6].

2. the *least presumptive explanation.* Explanation E_1 is less presumptive than E_2 if $F \cup E_2 \models E_1$. That is, if E_1 makes less (in terms of what can be implied) assumptions than E_2.

3. the *minimal abnormality explanation.* Explanation E_1 with conjecture assumptions P_1 and default assumptions D_1 is less abnormal than E_2 with assumptions $< P_2, D_2 >$ if $F \cup E_2 \models P_1$ and either $F \cup E_1 \not\models P_2$ or ($F \cup E_1 \models P_2$ and $F \cup E_2 \models D_1$). That is, if it makes less abnormality assumptions or it makes the same abnormality assumptions and fewer normality assumptions.

The first two can both be seen as preferring the minimal explanation. The first is the syntactic minimal explanation, where we treat a scenario as a set of axioms, and the second is the semantically minimal explanation where we equate a scenario with its logical theory (or its set of models).

The third definition is more heuristic and depends on how the domain is represented. It can be seen as a formulation of the maxim "if there is nothing wrong, don't fix it"; we don't even want to hypothesise errors unless there is evidence for them. There may, however, be a correct explanation, which is not one of the minimal abnormality explanations (see example 3.4).

I cannot think of a situation where one would not want the minimal explanation (i.e., why one would want to make extra unneeded assumptions).

[6] I am not advocating comparing scenarios by counting the number of assumptions in them. Such comparators have too many problems of slight changes to the representation of the problem domain giving different answers. For example it is not reasonable to always prefer one rare disease over two common diseases.

If there is a correct explanation, there is a minimal explanation which is also correct, as the following lemma indicates:

Lemma 3.2 *If there is an explanation true in an interpretation, there is a minimal explanation true in that interpretation.*

> **Proof:** Suppose explanation E of g is true in interpretation I. By the compactness theorem of the predicate calculus, there is a finite subset of E which is also an explanation of g. If we consider all of the subsets of E, one is a minimal explanation of g, and it is true in I. □

Thus, by restricting ourselves to the minimal explanations we will not remove the only correct explanation.

Although there are cases where no least presumptive explanation exists (example 3.7) as well as cases where it can be argued that the least presumptive explanation may not be the "best" explanation (example 3.8), it seems as though the least presumptive explanation is often the desired explanation.

Example 3.3 Let

$$\Pi = \{broken(leg), broken(tibia)\}$$

$$\Delta = \{broken(leg) \Rightarrow sore(leg)\}$$

$$F = \{broken(tibia) \Rightarrow broken(leg)\}$$

if we observe $sore(leg)$ there is one least presumptive explanation:

$$\{broken(leg), broken(leg) \Rightarrow sore(leg)\}$$

That is, we conjecture that the person has a broken leg and that the broken leg caused the sore leg. The explanation:

$$\{broken(tibia), broken(leg) \Rightarrow sore(leg)\}$$

is another minimal explanation, however it is not a least presumptive explanation. There is no evidence that the tibia is broken over the leg is broken; assuming the tibia is broken implies that the leg is broken.

Example 3.4 Consider the following system[7]:

$$\begin{aligned}
\Delta = \{\ & \textit{bird-so-flies}(X),\\
& \textit{emu-so-doesn't-fly}(X),\\
& \textit{flying-emu-so-flies}(X),\\
& \textit{bird-so-feathered}(X)\}\\
\Pi = \{\ & \textit{bird}(X),\\
& \textit{emu}(X),\\
& \textit{flyingemu}(X)\}\\
F = \{\ & \forall X\ \textit{bird}(X) \land \textit{bird-so-flies}(X) \Rightarrow \textit{flies}(X),\\
& \forall X\ \textit{emu}(X) \land \textit{emu-so-doesn't-fly}(X) \Rightarrow \neg \textit{flies}(X),\\
& \forall X\ \textit{flyingemu}(X) \land \textit{flying-emu-so-flies}(X) \Rightarrow \textit{flies}(X),\\
& \forall X\ \textit{emu}(X) \Rightarrow \textit{bird}(X),\\
& \forall X\ \textit{flyingemu}(X) \Rightarrow \textit{emu}(X),\\
& \forall X\ \textit{bird}(X) \land \textit{bird-so-feathered}(X) \Rightarrow \textit{feathered}(X)\\
& \forall X\ \textit{emu}(X) \Rightarrow \neg \textit{bird-so-flies}(X),\\
& \forall X\ \textit{flyingemu}(X) \Rightarrow \neg \textit{emu-so-doesn't- fly}(X)\}
\end{aligned}$$

If we observe that *Polly* is feathered, there is one least presumptive explanation, namely

$$\{\textit{bird}(\textit{polly}), \textit{bird-so-feathered}(\textit{polly})\}$$

There are other explanations for the observation, for example

$$\{\textit{emu}(\textit{polly}), \textit{bird-so-feathered}(\textit{polly}), \textit{flying-emu-so-flies}(\textit{randy})\}$$

but each of these makes extra assumptions for which there is no evidence (and, together with F, imply the least presumptive explanation).

If we observe that Tweety flies, there are two least presumptive explanations:

1. Tweety is a bird, and Tweety flies because birds fly. This is given by the explanation

$$\{\textit{bird}(\textit{tweety}), \textit{bird-so-flies}(\textit{tweety})\}$$

[7] Here we are using the technique of naming possible hypotheses [Poole88a].

2. Tweety is a flying emu, and Tweety flies, because flying emus, by default, fly. This is given by the explanation

$$\{flyingemu(tweety), flying\text{-}emu\text{-}so\text{-}flies(tweety)\}$$

The first explanation is the minimal abnormality explanation, as it makes less assumptions about Tweety than the second (as it only assumes Tweety is a bird, not that she is a flying emu). As far as we have evidence, either explanation could be correct; we do not want to make any abnormality assumptions for which we do not have evidence. We have evidence that Tweety is a bird, but we do not have the extra evidence that Tweety is a flying emu.

The following two theorems give relationships between the three comparators.

Theorem 3.5 *A least presumptive explanation is always logically equivalent to a minimal explanation.*

> **Proof:** Suppose E is a least presumptive explanation and suppose that E' is an explanation such that $E' \subset E$, then $E \models E'$, so $E' \models E$ otherwise E' is less presumptive than E. So if there is a smaller explanation than a least presumptive explanation, they are equivalent. □

That this does not mean that a least presumptive explanation (as defined) is always a minimal explanation. We can add hypotheses and conjectures implied by a least presumptive explanation to the explanation; it is still least presumptive, but no longer minimal. The above theorem shows that nothing is lost by assuming that all least presumptive explanations are minimal; in the rest of this paper this assumption is made.

Theorem 3.6 *A minimal abnormality explanation is always a least presumptive explanation.*

> **Proof:** Suppose E is a minimal abnormality explanation with assumptions $< P, D >$. We need to prove that there cannot be an explanation which is strictly less presumptive than E.

Assume that explanation E', with assumptions $< P', D' >$, is strictly less presumptive than E (i.e., $E \models E'$ and $E' \not\models E$); we want to show that E' is strictly less abnormal than E.

We know $E \models P'$ and $E \models D'$ (as $E \models E'$). $E' \not\models P$ or $E' \not\models D$ otherwise $E' \models P \wedge D$ and so $E' \models E$. So we know $E \models P'$ and ($E' \not\models P$ or $E' \not\models D$) and $E \models D'$, and so $E \models P'$ and $E' \not\models P$ or ($E' \not\models D$ and $E \models D'$), that is, E' is less abnormal than E.

Suppose E is less abnormal than E'. In this case $E' \models P$ and, as we know $E \models P'$, $E' \models D$. We then have $E' \models P \wedge D$ so $E' \models E$, a contradiction to E' being strictly less presumptive than E.

So if E is a minimal abnormality explanation, there is no strictly less presumptive explanation. □

Example 3.4 shows that the converse is not always true.

The following example shows that there is not always a least presumptive explanation:

Example 3.7 Consider the following system:

$$\begin{aligned}
\Pi &= \{p(X)\} \\
F &= \{\ \forall N\ p(N) \Rightarrow p(N+1), \\
&\qquad int(0), \\
&\qquad \forall N\ int(N) \Rightarrow int(N+1), \\
&\qquad \forall X\ (int(X) \wedge p(X) \Rightarrow g)\}
\end{aligned}$$

There is no least presumptive explanation of g, but rather an infinite chain of less presumptive explanations. There are infinitely many minimal explanations of g (one for each integer).

There are also cases where one can argue that the least presumptive explanation is not necessarily the best explanation:

Example 3.8 Suppose we are building a user modelling system, and want to be able to conjecture the interests of people and have the following conjectures:

$\Pi = \{$ *interested-in-hardware*,
interested-in-formal-AI,
interested-in-logic,
interested-in-CS$\}$

The defaults of the interests are given as defaults:

$\Delta = \{$ *interested-in-hardware* $\Rightarrow$ *interested-in-logic* $\wedge$ *interested-in-CS*,
interested-in-formal-AI $\Rightarrow$ *interested-in-logic* $\wedge$ *interested-in-CS*,
interested-in-logic $\Rightarrow$ *borrows-logic-books*,
interested-in-CS $\Rightarrow$ *writes-computer-programs*$\}$

If we observe that someone borrows logic books, it is reasonable to conjecture that they are interested in logic. This is the least presumptive explanation. If we observe that someone borrows logic books and writes computer programs, there is one least presumptive explanation, namely that they are interested in computer science and interested in logic. The alternate explanations, namely that they are interested in formal AI or interested in hardware are not going to be least presumptive, although one could argue that they are the best explanations on the grounds of simplicity. The disjunct of instances of a general law is always less presumptive than the general law, although it could be argued that the general law is a better explanation. It may be better to get to the root cause of a problem than to just give the weakest explanation.

This is similar to what was argued in [Popper62, p. 219] that one does not always want the most likely explanation (the most likely always being least presumptive).

Some work has been done on defining appropriate scenario comparators. [Popper62] proposes a *verisimilitude* for comparing theories and [Quine78, chapter 6] defined five *virtues* on which to compare explanations. [Poole85], [Goebel87] and [Neufeld87] define different scenario comparators. Much more work needs to be done in this area.

4 A Default and Abductive Reasoning System

The architecture we are considering is one where the system is provided with facts, defaults and conjectures. We assume these provide the general knowledge about the domain being modelled (e.g., how diseases interact and how symptoms work in a diagnosis system, and general knowledge about objects, occlusion etc., in a recognition task). All specific knowledge about a particular case is added as observations.

A sequence of observations is provided to the system. The system constructs the best (according to the explanation comparisons given) explanations of the observations. From each explanation we can ask what is predicted. The system can also propose what observations it would like about the world in order to prune and refine its explanations.

4.1 Interacting with the system

When implementing Theorist we want a system in which we can add facts, defaults, etc., and give observations and ask predictions based on what the system has been told.

The state of the system can be described as a tuple

$$< F, \Delta, \Pi, O, \mathcal{E} >$$

where

F is the set of facts

Δ is the set of defaults

Π is the set of conjectures

O is the set of observations that have been made

$\mathcal{E}$ is the set of preferred (according to some preference criteria) explanations of the observations O.

The input language to the system is defined below. The syntax of each command is given, along with how the command affects the state of the system, assuming the current state is $< F, \Delta, \Pi, O, \mathcal{E} >$.

fact w.

where w is a formula, means "$\forall w$"[8] is a new fact. The resulting state is

$$< F \cup \{\forall w\}, \Delta, \Pi, O, \mathcal{E}' >$$

where $\mathcal{E}'$ is the resulting explanations given $\forall w$ as a fact (section 6).

default n.

where n is a name (predicate with only free variables as arguments) means n is a new default. Formally this means that the new state is

$$< F, \Delta \cup \{n\}, \Pi, O, \mathcal{E}' >$$

where $\mathcal{E}'$ is the resulting explanations given the new default.

default $n : w$.

where w is a formula, and n is a name, means that w is a default, with name n[9]. The new state is

$$< F \cup \{\forall(n \Rightarrow w)\}, \Delta \cup \{n\}, \Pi, O, \mathcal{E}' >$$

conjecture n.

where n is a name means that n is a new conjecture. The new state is

$$< F, \Delta, \Pi \cup \{n\}, O, \mathcal{E}' >$$

conjecture $n : w$.

where w is a formula, and n is a name, means w is a formula with name n. The new state is

$$< F \cup \{\forall(n \Rightarrow w)\}, \Delta, \Pi \cup \{n\}, O, \mathcal{E}' >$$

[8] $\forall w$ is the universal closure of w, that is, if w has free variables $\overline{v}$ then $\forall w$ means $\forall \overline{v}\ w$. Similarly $\exists w$ is the existential closure of w.

[9] See [Poole88a] for a discussion on naming defaults.

observe g.

where g is a closed formula, means that g is a new observation. The new $\mathcal{E}$ is the set of preferred explanations of all of the observations (i.e., $O \wedge g$).

predict g, S.

where g is a formula and S is a scenario (usually one of the elements of $\mathcal{E}$), returns *yes* (together with the instance) if some instance of g is in every extension of S and *no* otherwise.

predict g.

where g is a formula returns *yes* (together with the instance) if some instance of g is in every extension of E, Δ for all $E \in \mathcal{E}$, and *no* otherwise.

For prediction, if the answer is *yes*, the set of explanations of g for which there is no mutually inconsistent scenario (the set $\mathcal{E}$ of theorem 2.6) is returned. If the answer is *no*, the scenario from which g cannot be explained (the set S of point 3 of theorem 2.6) is returned. Note that the answer "no" does not mean we predict g is false, but rather we do not predict g is true.

Example 4.1 Consider the following example:

A person can possibly have a brain tumour,
a person can possibly have a broken leg,
a brain tumour typically produces a head ache, and
a broken leg typically produces a sore leg and a bent leg.

This knowledge can be represented as:

conjecture *brain-tumour.*
conjecture *broken-leg.*
default *tumoured-heads-ache: brain-tumour* ⇒ *head- ache.*
default *broken-legs-are-sore: broken-leg* ⇒ *sore-leg.*
default *broken-legs-are-bent: broken-leg* ⇒ *bent-leg.*

If we make the observation

observe *bent-leg.*

we have one minimal and least presumptive explanation:

{*broken-leg, broken-legs-are-bent*}

If we subsequently ask:

predict *head-ache.*

the answer is *no* (it cannot be explained). If we ask

predict *sore-leg.*

the answer is *yes*; the returned explanation is

{*broken-legs-are-sore*}

Example 4.2 (Pearl) [Pearl87, p. 371] gives the following example to argue that there should be a distinction between *causal rules* and *evidential rules.* Here we show how the problems he was trying to solve do not arise in our system. We add the causal rules as defaults (or facts if we do not want to consider them having exceptions)

default *rained-so-wet: rained-last-night* ⇒ *grass-is-wet.*
default *sprinkled-so-wet: sprinkler-was-on* ⇒ *grass-is-wet.*
default *wet-so-cold: grass-is-wet* ⇒ *grass-is-cold-and-shiny.*
default *grass-wet-so-shoes-wet: grass-is-wet* ⇒ *shoes-are-wet.*

Instead of adding the reverse of these rules as evidential rules [Pearl87], we make the possible causes we are considering as conjectures:

conjecture *rained-last-night.*
conjecture *sprinkler-was-on.*

If we observe that it rained last night, we have one explanation:

{*rained-last-night*}

From this we can predict that the grass is wet, that the grass is cold and shiny and that my shoes are wet. There is no way to predict that the sprinkler was on last night (which was the problem with having the evidential rules as explicit rules).

If we had instead observed that the grass is cold and shiny, there are two explanations:

$$\{\textit{rained-last-night, rained-so-wet, wet-so-cold}\}$$

$$\{\textit{sprinkler-was-on, sprinkled-so-wet, wet-so-cold}\}$$

From both of these we predict that my shoes are wet.

5 Implementation

In this section we show how a theorem prover (see e.g., [Chang73]) can be used to implement this system.

One of the things that is important is whether we can localise search rather than always having to do a full consistency check. We prefer to search only that part of the space relevant to what is being added or asked; we would like to know when parts of the knowledge base are irrelevant.

One way that this can be done is to assume only a limited form of completeness of the theorem prover. We want our theorem prover to be sound, but only require completeness in the sense that if there is a relevant proof of some goal, it can be found. A proof of g from A (denoted $A \vdash g$) is assumed to be sound (i.e., if $A \vdash g$ then $A \models g$), but it need only be complete in the sense that if A is consistent and $A \models g$ then $A \vdash g$. Linear Resolution [Chang73] with head clause g is such a proof procedure. Such deduction systems can be more efficiently implemented than complete theorem provers as they do not need to consider irrelevant reasons for something following from a set of axioms.

5.1 Explanation

The following two theorems are important for implementing the system.

Theorem 5.1 *If A is consistent, g is explainable from A, H if and only if there is a ground proof of g from $A \cup D$ where $D = \{d_1, ..., d_n\}$ is a set of ground instances of elements of H such that $A \wedge \{d_1, ..., d_{i-1}\} \not\vdash \neg d_i$ for all $i = 1..n$.*

Proof: If g is explainable from A, H, there is a set D of ground instances of elements of H such that $A \cup D \models g$ and $A \cup D$ is consistent, so there is a proof of g from $A \cup D$. $A \cup D$ is consistent so there can be no sound proof of inconsistency. That is, we cannot prove $A \wedge \{d_1, ..., d_{i-1}\} \vdash \neg d_i$ for any i.

If there is a proof of g from $A \cup D$ then $A \cup D \models g$. If $A \cup D$ is inconsistent there is some least i such that $A \cup \{d_1, ..., d_i\}$ is inconsistent. We know $A \cup \{d_1, ..., d_{i-1}\}$ is consistent and $A \cup \{d_1, ..., d_{i-1}\} \models \neg d_i$ so $A \cup \{d_1, ..., d_{i-1}\} \vdash \neg d_i$. So, if there is no i such that $A \cup \{d_1, ..., d_{i-1}\} \vdash \neg d_i$ then $A \cup D$ is consistent. □

This leads us to the following algorithm to explain g from A, H:

1. Try to prove g from $A \cup H$; make D the set of instances of elements of H used in the proof.

2. Reject D if it contains a Skolem function. This is enforcing the groundedness of explanations[10].

3. Ground D (substitute a new constant for each of the free variables in D)[11]. We thus have created a ground proof of g from $A \cup D$.

4. For each $d_i \in D$, try to prove $\neg d_i$ from $A \wedge \{d_1, ..., d_{i-1}\}$. If all such proofs fail, D is an explanation for g.

[Poole88b] gives the details of how explanation can be implemented by compiling Theorist into Prolog; [PGA87] gives a Prolog interpreter for explanation.

There is a strong resemblance between this algorithm and negation as failure [Clark78]. We conclude hypotheses by failing to prove their negations. Apart from the more powerful logic (disjunction and explicit negation) used here, the main difference is that we fail to prove the negation

[10] See [Poole87a] for a discussion about relaxing the groundedness of scenarios.

[11] This is correct whether we interpret the free variables as universally quantified, or as schema denoting each individual (as discussed in section 3.1). In the former case this grounding is Skolemisation [Chang73], in the latter case this is just choosing an individual to assume. We will only be able to show inconsistency if we could show inconsistency for any instance.

in a simpler system than the top level system. Instead of failing to *explain* the negation of a hypothesis, we fail to *prove* the negation of a hypothesis from the facts and the previously assumed hypotheses. One advantage of Theorist is that in a decidable logic (e.g, the propositional calculus), explainability is also decidable. This is not the case for negation as failure (consider the formula $p \leftarrow \neg p$).

5.2 Prediction

Consider the question of whether some proposition is in all extensions (the other cases of prediction are straightforward to implement given the previous section).

The naive way to do this (generating extensions and testing membership) does not work for two reasons:

1. extensions are infinite. Even if we consider the generators of the extensions (i.e., the maximal scenarios), we still get the same problem as these are also usually infinite.

2. there are potentially an infinite number of extensions.

Is there a way to implement this so that we only need to consider the relevant parts of the relevant extensions? What are the relevant parts and the relevant extensions needed to determine that g is in all extensions? This section provides answers to these questions.

Point 4 of theorem 2.6 leads to the following dialectical view of membership in every extension.

There are two processes $\mathcal{Y}$ and $\mathcal{N}$ that are having an argument as to whether g should be predicted. Process $\mathcal{Y}$ tries to find explanations of g. Process $\mathcal{N}$ tries to find a scenario inconsistent with all of $\mathcal{Y}$'s explanations.

First $\mathcal{Y}$ tries to find an explanation D of g. Then $\mathcal{N}$ tries to find a scenario inconsistent with D (i.e., an explanation of $\neg D$). $\mathcal{Y}$ must then try to explain g given $\mathcal{N}$'s scenario.

In general $\mathcal{Y}$ has a set of explanations Φ. $\mathcal{N}$ tries to find a scenario S which is inconsistent with all members of Φ (i.e., explains the conjunction of the negation of the elements of Φ). When $\mathcal{N}$ finds scenario S, $\mathcal{Y}$ must find an explanation of g from S. Whichever process, using a complete (in the sense of section 5) proof procedure, gives up first loses:

- If $\mathcal{Y}$ cannot come up with an explanation based on $\mathcal{N}$'s scenario S, then g is not in all extensions (in particular g is not in any extension of S).
- If $\mathcal{N}$ cannot come up with a scenario inconsistent with all of $\mathcal{Y}$'s arguments, every extension contains at least one of $\mathcal{Y}$'s arguments, and so g is in every extension.

Example 5.2 Consider example 2.1, and the process of trying to determine *pro-star-wars(dick)*. We have the following dialogue:

$\mathcal{Y}$: $\mathit{republican}(\mathit{dick}) \Rightarrow \mathit{hawk}(\mathit{dick}), \mathit{hawk}(\mathit{dick}) \Rightarrow \mathit{pro\text{-}star\text{-}wars}(\mathit{dick})$
$\mathcal{N}$: $\mathit{quaker}(\mathit{dick}) \Rightarrow \mathit{dove}(\mathit{dick})$
$\mathcal{Y}$: no explanation

$\mathcal{Y}$ can find no explanation of *pro-star-wars(dick)* from the scenario given by $\mathcal{N}$. Thus, we do not conclude *pro-star-wars(dick)*.

Consider the process of determining *politically-motivated(dick)*:

$\mathcal{Y}$: $\mathit{quaker}(\mathit{dick}) \Rightarrow \mathit{dove}(\mathit{dick}), \mathit{dove}(\mathit{dick}) \Rightarrow \mathit{politically\text{-}motivated}(\mathit{dick})$
$\mathcal{N}$: $\mathit{republican}(\mathit{dick}) \Rightarrow \mathit{hawk}(\mathit{dick})$
$\mathcal{Y}$: $\mathit{republican}(\mathit{dick}) \Rightarrow \mathit{hawk}(\mathit{dick}), \mathit{hawk}(\mathit{dick}) \Rightarrow \mathit{politically\text{-}motivated}(\mathit{dick})$
$\mathcal{N}$: no explanation

We conclude politically-motivated(dick).

There are a few points to notice about this algorithm.

1. $\mathcal{Y}$'s explanations of g from the S's generated by $\mathcal{N}$ are explanations of g from A, H. Thus we can implement $\mathcal{Y}$ as finding successive explanations of g from A, H. We do not need to start from scratch when $\mathcal{N}$ has found a contradictory scenario, but can just continue generating explanations. $\mathcal{N}$'s explanations can be used to prune this search, as any partial explanation that has already been shown to be inconsistent with a scenario generated by $\mathcal{N}$ can be pruned.

2. $\mathcal{N}$ also does not need to start from scratch each time $\mathcal{Y}$ generates a new explanation of g. Suppose $D_1, ..., D_{n+1}$ are the explanations generated by $\mathcal{Y}$. E_{n+1} is an explanation of $\neg D_1 \wedge ... \wedge \neg D_{n+1}$ if and only if there is some E_n, an explanation of $\neg D_1 \wedge ... \wedge \neg D_n$, such that E_{n+1} is E_n together with an explanation of $\neg D_{n+1}$ from $F \cup E_n$. This implies that $\mathcal{N}$ can generate the new explanations from the old explanations, and the newly generated goal.

   If the set of all explanations is maintained, this procedure is very much like a non-propositional, non-Horn ATMS [de Kleer86]. Both space considerations and the desire to do as little redundant work as necessary, would probably support the alternative of maintaining one search tree; each time $\mathcal{Y}$ comes up with a new explanation, $\mathcal{N}$ continues the search to prove the negation of that goal. $\mathcal{N}$ does not need to redo the work to find an explanation of the old explanations. $\mathcal{N}$ may, however, need to find alternate proofs of the old explanations.

Note that the set of explanations referred to in point 4 of theorem 2.6 is countable, but not necessarily finite. The following example has an infinite set of possible explanations to check. The preceding algorithm will not halt on this example.

Example 5.3 Consider

$$\begin{aligned} H &= \{\ p(X)\} \\ F &= \{\ q(0), \\ &\qquad \forall N\ q(N) \Rightarrow q(s(N)), \\ &\qquad pos(s(0)), \\ &\qquad \forall N\ pos(N) \Rightarrow pos(s(N)), \\ &\qquad \forall N\ pos(N) \Rightarrow lt(0,N), \\ &\qquad \forall N \forall M\ lt(M,N) \Rightarrow lt(s(M),s(N)), \\ &\qquad \forall N \forall M\ lt(M,N) \Rightarrow \neg(p(M) \wedge p(N)), \\ &\qquad (\exists X\ p(X) \wedge q(X)) \Rightarrow g\} \end{aligned}$$

q is true of all non-negative integers, and p is true of at most one non-negative integer. There are infinitely many extensions, one for each positive integer (each one containing $p(n)$ for some positive integer n). g is in all extensions, but there is no finite set of proofs which are applicable for all extensions, without jumping out of the system and arguing as we have done here.

6 Building and Maintaining the Knowledge Base

There are a number of choices that the designer of a system can make as to how the knowledge base is maintained. The following are possible:

1. record just what was explicitly told and compute all answers when asked.

2. maintain one explanation for the observations and build another if this one proves to be wrong (e.g., [Doyle79]).

3. maintaining multiple, but not all explanations. For example, maintaining just those minimal abnormality explanations and only considering others if these prove inadequate. As example 6.4 below shows, it is often difficult to ensure that one is maintaining the minimal abnormality explanations without also maintaining all of the other least presumptive explanations.

4. maintaining parts of all of the least presumptive explanations. This may make it easier to see when one explanation can be replaced by a better explanation. For example [Neufeld87] describes an algorithm which always maintains the most likely explanation by maintaining enough of other explanations to ensure that they will be less likely than the preferred one.

5. maintain all least presumptive explanations (or all minimal explanations). This algorithm would correspond to a non-propositional, non-Horn ATMS [de Kleer86].

6. maintaining a representation of all extensions (e.g., the generating hypotheses). This may make building the knowledge base inefficient, but may make it easier to query.

Which of these is better may depend on efficiency grounds (minimising space, time or interaction with the user) as well as psychological grounds (e.g., wanting to model an agent who follows one line of belief and only changes their mind when they are forced to, or an agent that doesn't consider some line of reasoning unless other lines have been exhausted).

If we maintain explanations we do not want to recompute everything after each input. In the next sections we consider how adding facts, defaults, hypotheses and observations affects the explanations generated.

6.1 Incremental Observations

One of the things that would be nice to know is to what extent one can incrementally build explanations for observations as they come in. We are assuming that we do not just receive one big conjunction of all observations, but rather get our observations incrementally. We would like to know that the explanations built incrementally are the same as those built from the conjunction of the observations. In this section we show that this is the case if we maintain minimal explanations or least presumptive explanations, but not if we just maintain minimal abnormality explanations.

Theorem 6.1 *We can build minimal explanations incrementally:*
If $S_1, ..., S_n$ are the minimal explanations of g_1 from $(F, \Pi \cup \Delta)$ then the minimal elements of the set of explanations of g_2 from $(S_i, \Pi \cup \Delta)$ for some S_i, are exactly the minimal explanations of $g_1 \wedge g_2$ from $(F, \Pi \cup \Delta)$.

> **Proof:** If E is an explanation of $g_1 \wedge g_2$ from F, Π, Δ then E is an explanation of g_1 from F, Π, Δ, so there is some $S \subseteq E$ such that S is a minimal explanation of g_1. Then E is an explanation of g_2 from S, and is minimal.
>
> Similarly if E is an explanation of g_2 from some S_i, E is an explanation of $g_1 \wedge g_2$ from F. Hence, the minimal explanations of g_2 from the S_i are the minimal explanations of $g_1 \wedge g_2$ from F. □

Theorem 6.2 *If $S_1, ..., S_n$ are the least presumptive explanations for g_1 from F, Π, Δ, the following are equivalent*

1. *S is a least presumptive explanation of $g_1 \wedge g_2$ from F, Π, Δ.*
2. *S is a least presumptive scenario of the explanations of g_2 from S_i, Π, Δ. That is, it is a minimal element, in terms of least presumptiveness, of the set $\{E : E \text{ is an explanation of } g_2 \text{ from } S_i, \Pi, \Delta \text{ for some } i\}$.*

Proof: 1 ⇒ 2. Suppose S is a least presumptive explanation of $g_1 \wedge g_2$ from F. S is an explanation of g_1, so one S_i implies S. S is an explanation of g_2 from S_i; we need to show that it is least presumptive. Suppose S' is a strictly less presumptive explanation of g_2 from S_i, then it is an explanation of $g_1 \wedge g_2$ from F less presumptive than S, a contradiction to the minimality of S.

2 ⇒ 1. Suppose S is a least presumptive explanation of g_2 from S_i. S is an explanation of $g_1 \wedge g_2$ from F. We need to show that S is least presumptive. If S' is a strictly less presumptive explanation of $g_1 \wedge g_2$ from F, it is also an explanation of g_1 from F, so there is some S_i which implies it (by the minimality of the S_i). S' is an explanation of g_2 from S_i, which is less presumptive than S, a contradiction to the minimality of S, so no such S' can exist. □

This leads us to a way to think about the system, namely that there is a sequence of observations, and we collect all the minimal or least presumptive theories at each step. At the end of the observations, we know we have the least presumptive explanations for the conjunction of the observations.

These theorems do not mean that we can build explanations in isolation of each other, without considering the other (minimal or least presumptive) explanations. Consider the following example

Example 6.3 Let

$$\begin{aligned}
&\Pi = \{a, b, c\} \\
&\Delta = \{\} \\
&F = \{\ a \Rightarrow g_1, \\
&\qquad b \Rightarrow g_1 \wedge g_2\}
\end{aligned}$$

If we observe g_1 there are two minimal (and least presumptive) explanations: $\{a\}$ and $\{b\}$. If we subsequently observe g_2, there is one minimal explanation, namely $\{b\}$. We can explain g_2 from $\{a\}$, (using the explanation $\{a, b\}$) but this explanation is subsumed by a simpler explanation from $\{b\}$.

Theorem 6.2 does not work for minimal abnormality explanations. Consider the following example:

Example 6.4 Let

$$\begin{aligned}
&\Pi = \{a, b, c\}\\
&\Delta = \{d_1, d_2, d_3\}\\
&F = \{\ a \wedge b \wedge d_1 \Rightarrow g_1 \wedge g_2,\\
&\qquad a \wedge d_2 \Rightarrow g_1,\\
&\qquad b \wedge c \wedge d_3 \Rightarrow g_2\}
\end{aligned}$$

The least presumptive explanations for g_1 are

$$\{a, b, d_1\}$$

$$\{a, d_2\}$$

the second of which is the minimal abnormality explanation. The least presumptive explanations for $g_1 \wedge g_2$ are

$$\{a, b, d_1\}$$

$$\{a, d_2, b, c, d_3\}$$

the first of which is the minimal abnormality explanation.

This means that we cannot simply find the minimal abnormality explanation by maintaining minimal abnormality explanations and using them to explain new observations.

6.2 Adding new facts

In this section we wish to answer the question of how the set of explanations should be changed when a new facts is added. A new fact may remove old explanations (by making them inconsistent or making one explanation less presumptive than a previously least presumptive explanation) or add new explanations.

The command

fact w.

means that the knowledge base is changed from

$$< F, \Delta, \Pi, O, \mathcal{E} >$$

to

$$< F \cup \{\forall w\}, \Delta, \Pi, O, \mathcal{E}' >$$

We would like to know how the set of explanations has changed by adding this new fact. We would like to build the new $\mathcal{E}'$ from the old $\mathcal{E}$ by only doing local search from the newly added fact. In general we would like to build $\mathcal{E}'$ by adding and removing elements from $\mathcal{E}$.

For all $E \in \mathcal{E}$ we know

$F \cup E \models O$
$F \cup E$ is consistent.

If $E' \in \mathcal{E}'$ then $F \cup \{\forall w\} \cup E' \models O$ so either

1. $F \cup E' \models O$ in which case E' is an explanation of O from F. $E' \in \mathcal{E}$ as there can be no smaller explanation of O from F, otherwise it is a smaller explanation of O from $F \cup \{\forall w\}$. We can thus carry over the old explanation from $\mathcal{E}$.

2. $F \cup E' \not\models O$ and so $F \cup E' \cup \neg O$ is consistent and implies $\neg \forall w$. This is the only case where we will add explanations to $\mathcal{E}$.

The newly added fact may make some previous explanations inconsistent. Suppose $E \in \mathcal{E}$; E is not in $\mathcal{E}'$ if $F \cup \{\forall w\} \cup E$ is inconsistent. In this case $F \cup E$ is consistent and implies $\neg \forall w$, and so there is a proof of $\neg \forall w$ from $F \cup E$.

This implies that when a new fact is added, we need to do three things:

1. try to explain $\neg\forall w$ from $F \cup \neg O, \Delta \cup \Pi$. The generated explanation should be checked consistent with $F \cup \{\forall w\}$. Each explanation should be added to $\mathcal{E}$.

2. try to prove $\neg\forall w$ from $F \cup E$, for each $E \in \mathcal{E}$, and remove any explanation which is proven inconsistent.

3. remove any explanations which are no longer minimal (as the first step may have created an explanation simpler than a previous explanation).

For each of these steps we only need to do a local search from the newly added fact.

If we maintain least presumptive explanations, we have to consider that the newly added fact may make one explanation which was previously least presumptive no longer least presumptive. This can happen by the newly added fact adding an implication between two previously least presumptive explanations. Suppose E' is less presumptive than E when $\forall w$ is a fact and is not otherwise. That is $F \cup \{\forall w\} \cup E \models E'$ and $F \cup E \not\models E'$ and so $\neg\forall w$ can be proven from consistent $F \cup E \cup \neg E'$. This can be recognised by trying to explain $\neg\forall w$ from $F \cup E, \Delta \cup \Pi$ for each $E \in \mathcal{E}$.

6.2.1 Adding Defaults and Conjectures

Consider the problem of adding the default

default $d : w$.

where d is a new name (as we would normally expect it to be). Note that exactly the same analysis carries through for adding conjectures.

Theorem 6.5 (Semimonotonicity) *If $\mathcal{E}$ is the set of explanations before the default was added and $\mathcal{E}'$ the explanations after, then $\mathcal{E} \subseteq \mathcal{E}'$.*

> **Proof:** If $E \in \mathcal{E}$ then $F \cup E \models O$ and so $F \cup \{\forall d \Rightarrow w\} \cup E \models O$. $F \cup E$ is consistent, and so has a model M. The model which is the same as M but with all instances of d false is a model for $F \cup \{\forall d \Rightarrow w\} \cup E$. So E is an explanation of O from

> $F \cup \{\forall d \Rightarrow w\}, \Delta \cup d, \Pi$. It is minimal as any smaller explanation would also be an explanation of O from F, Δ, Π, as "$\forall d \Rightarrow w$" cannot play a role if d does not appear in E, F, O, Δ or Π. □

We now have to consider the case of there being a new explanation of O by virtue of the default being added. Suppose $E \in \mathcal{E}' - \mathcal{E}$. We know

$$F \cup \{\forall d \Rightarrow w\} \cup E \models O$$

There is some instance δ of d in E (otherwise $E \in \mathcal{E}$). $F \cup \{\forall d \Rightarrow w\} \cup (E - \{\delta\}) \cup \{\neg O\}$ is consistent (otherwise E is not minimal) and implies $\neg\delta$.

Hence when a new default is added we need to try to explain $\neg d$ from $F \cup \{\forall d \Rightarrow w\} \cup \{\neg O\}, \Delta \cup \{d\} \cup \Pi$, checking consistency with $F \cup \{\forall d \Rightarrow w\}$.

6.3 Detecting Multiple Extensions

In section 2.4 it was argued that one reasonable way to handle multiple extensions is to regard them as a bug that must be fixed up. What is needed is a way to detect when we have multiple extensions.

Suppose we have given A (these can be the facts or any other scenario we are interested in) and hypotheses H. As facts or hypotheses are added, the following theorems show how we can detect multiple extensions.

Theorem 6.6 Suppose (A, H) has one extension; $(A \cup f, H)$ has one extension if and only if whenever $\neg f$ is explainable from (A, H) by an explanation with more than one default, there is a subset of that explanation containing one default which is also an explanation of $\neg f$.

> **Proof:** Suppose E is a minimal explanation of $\neg f$ with more than one element. Choose $h \in E$ and let $E' = E - \{h\}$. We know $A \cup \{f\} \cup E'$ is consistent (by minimality of E), and $A \cup \{f\} \cup \{h\}$ is consistent, (by the minimality of E), but they are mutually inconsistent (as $A \cup \{f\} \cup E$ is inconsistent). They can be extended to different extensions.
>
> Conversely suppose $A \cup \{f\}$ has two extensions. Let E_1 and E_2 be the maximal sets of assumptions in each. $A \cup E_1 \cup E_2$ is

consistent (as A has only one extension). $A \cup E_1 \cup E_2 \cup \{f\}$ is inconsistent, as two extensions are always mutually inconsistent, so

$$A \cup E_1 \cup E_2 \models \neg f$$

by the compactness theorem of the first order predicate calculus, there are finite subsets S_1 and S_2 of E_1 and E_2 respectively such that

$$A \cup S_1 \cup S_2 \models \neg f$$

$A \cup S_1 \cup \{f\}$ is consistent (as it is a subset of an extension, so $S_2 \neq \{\}$. Similarly $S_1 \neq \{\}$. Thus there is an explanation of $\neg f$, namely $S_1 \cup S_2$, for which there is no one element subset that is an explanation of $\neg f$. □

Theorem 6.7 Suppose (A, H) has one extension; $(A, H \cup \{d\})$ has multiple extensions if and only if there is an instance d' of d, such that d' is consistent with A, and $\neg d'$ is explainable from (A, H).

Proof: Suppose $(A, H \cup \{d\})$ has multiple extensions. Suppose E_1 and E_2 are different extensions, then there are minimal sets of defaults $S_1 \subset E_1$ and $S_2 \subset E_2$ such that $A \cup S_1 \cup S_2$ is inconsistent. Neither S_i is empty, as the other is consistent with A. An instance d' of d must be in at least one of the S_i as A, H has only one extension. So d' is consistent with A, and $(S_1 \cup S_2) - \{d'\}$ is an explanation of $\neg d'$.

Conversely suppose d' is consistent with A and $\neg d'$ is explainable from A, H. Then there is an explanation E of $\neg d'$. $A \cup E$ and $A \cup d'$ are both scenarios and are mutually inconsistent, so can be extended to different extensions. □

These two theorems give a straightforward way to automatically detect multiple extensions.

7 Conclusion

In this paper I presented an architecture for both explaining observations and for making predictions. For each of these a number of possible definitions was discussed and compared. It seems as though no definition is correct for all situations; this paper is an attempt to compare different notions of each. An implementation was outlined which follows the semantics of minimal explanations and prediction being membership in all extensions.

One problem with this, is that all of the "algorithms" are undecidable in the worst case; they are not guaranteed to halt. In our, albeit limited, experience this has not been a problem. By using our system, we are learning how to "program" the logic to give us answers quickly. This is the topic of another paper, however.

An important feature of this work is that I have not proposed a new logic. I have tried to be careful in arguing that there are useful ways to use logic and have considered the consequences on building AI programs.

Acknowledgements

This work could not have been done without the ideas, criticism, feedback and support of Randy Goebel, Eric Neufeld, Bruce Kirby, Paul Van Arragon, Romas Aleliunas, Scott Goodwin and Bruce Spencer. Thanks to George Tsiknis, Alex Kean and Greg Provan for valuable comments on a previous version of this paper. This research was supported under NSERC grant OPPOO44121.

References

[Bobrow80] D. Bobrow, "Special Issue on Nonmonotonic Reasoning" *Artificial Intelligence*, Vol. 13.

[Cebulka88] K. d. Cebulka, S. Carberry and D. L. Chester, "Solving Dynamic-Input Interpretation problems using the hypothesise-test-revise paradigm", *Proc. fourth IEEE Conference on Artificial Intelligence Applications.*

[Chang73] C-L. Chang and R. C-T. Lee, *Symbolic Logic and Mechanical Theorem Proving*, Academic Press, 1973.

[Charniak85] E. Charniak and D. McDermott, *Introduction to Artificial Intelligence*, Addison-Wesley.

[Clark78] K. L. Clark, "Negation as Failure", in H. Gallaire and J. Minker (eds), *Logic and Data Bases*, pp. 119-140.

[Console89] L. Console and P. Torasso, "Hypothetical Reasoning in Causal Models", to appear *International Journal of Intelligent Systems.*

[Cox87] P. T. Cox and T. Pietrzykowski, *General Diagnosis by Abductive Inference*, Technical report CS8701, School of Computer Science, Technical University of Nova Scotia, April 1987.

[de Kleer86] J. de Kleer, "An Assumption-based TMS", *Artificial Intelligence, Vol. 28, No. 2*, pp. 127-162.

[Doyle79] J. Doyle, "A Truth Maintenance System", *Artificial Intelligence*, Vol. 12, pp 231-273.

[Edwards67] P. Edwards (ed.), *The Encyclopedia of Philosophy,* Macmillan, N.Y.

[Enderton72] H. B. Enderton, *A Mathematical Introduction to Logic*, Academic Press, Orlando.

[Genesereth87] M. R. Genesereth and N. Nilsson, *Logical Foundations of Artificial Intelligence*, Morgan–Kaufmann.

[Goebel87] R. G. Goebel and S. D. Goodwin, "Applying theory formation to the planning problem" in F. M. Brown (Ed.), *Proceedings of the 1987 Workshop on The Frame Problem in Artificial Intelligence*, Morgan Kaufmann, pp. 207-232.

[Hayes77] P. J. Hayes, "In Defence of Logic", *Proc. IJCAI-77*, pp. 559-565.

[Josephson87] J. R. Josephson, B. Chandrasekaran, J. R. Smith Jr., M. C. Tanner, "A Mechanism for Forming Composite Explanatory Hypotheses", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-17, No. 3, pp. 445-454.

[Kautz87] H. A. Kautz, "A Formal Theory for Plan Recognition", Department of Computer Science, University of Rochester, Technical Report TR-215.

[Moore82] R. C. Moore, "The Role of Logic in Knowledge Representation and Commonsense Reasoning", *Proc. AAAI-82*, pp. 428-433.

[Moore85] R. C. Moore, Semantical Considerations on Nonmonotonic Logic, *Artificial Intelligence*, Vol 25, No 1, pp. 75-94.

[Neufeld87] E. M. Neufeld and D. Poole, "Towards solving the multiple extension problem: combining defaults and probabilities", *Proc. AAAI Workshop on Reasoning with Uncertainty*, Seattle, July 1987.

[Pearl87] J. Pearl, "Embracing Causality in Formal Reasoning", *Proc. AAAI-87*, pp. 369-373.

[Poole85] D. L. Poole, "On the Comparison of Theories: Preferring the Most Specific Explanation", *Proc. IJCAI-85*, pp.144-147.

[PGA87] D. L. Poole, R. G. Goebel, and R. Aleliunas, "Theorist: a logical reasoning system for defaults and diagnosis", in N. Cercone and G.McCalla (Eds.) *The Knowledge Frontier: Essays in the Representation of Knowledge*, Springer Varlag, New York, 1987, pp. 331-352.

[Poole87a] D. L. Poole, "Variables in Hypotheses", *Proc. IJCAI-87*, pp. 905-908.

[Poole87b] D. L. Poole (Ed.), *Experiments in the Theorist Paradigm: A Collection of student papers on the Theorist Project*, Research Report CS-87-30, Department of Computer Science, University of Waterloo, May.

[Poole88a] D. L. Poole, "A Logical Framework for Default Reasoning", to appear *Artificial Intelligence*, Vol. 36, No. 1, pp. 27-47.

[Poole88b] D. L. Poole, *Compiling a Default Reasoning System into Prolog*, Research Report CS-88-01, Department of Computer Science, University of Waterloo.

[Poole88c] D. Poole, "Representing Knowledge for Logic-based Diagnosis", *Proceedings International Conference on Fifth Generation Computing Systems (FGCS-88)*, pp.??.

[Poole89] D. Poole, "What the lottery paradox tells us about default reasoning", to appear *First International Conference on the Principles of Knowledge Representation and Reasoning*, Toronto, May 1989.

[Popl73] H. Popl, "On the mechanisation of Abductive Logic", *Proc. IJCAI-73*, pp. 147-152.

[Popper62] K. R. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge*, Basic Books, New York.

[Quine78] W. V. Quine, and J. S. Ullian, *The Web of Belief*, Random House, Yew York, Second Edition.

[Reggia83] J. A. Reggia, D. S. Nau and P. Y. Wang, "Diagnostic expert systems based on a set covering model", *International Journal of Man-Machine Studies 19*, pp. 437-460.

[Reiter80] R. Reiter, "A Logic for Default Reasoning", *Artificial Intelligence*, Vol. 13, pp 81-132.