PROCEEDINGS OF THE INTERNATIONAL CONFERENCE
ON FIFTH GENERATION COMPUTER SYSTEMS 1992,
edited by ICOT. © ICOT, 1992

Hypothetico-deductive Reasoning

Chris Evans[*] and Antonios C. Kakas[†]

[*]Department of Mathematical Studies, Goldsmiths' College, University of London
New Cross, London SE14 6NW, UK. EMAIL: c.evans@gold.lon.ac.uk.

[†]Department of Computer Science, University of Cyprus, 75 Kallipoleos Street,
Nicosia, Cyprus. EMAIL: kakas@cyearn.earn

(Part of the research for this paper was completed while both authors were at Imperial College, London SW7 2BZ)

Abstract

This paper presents a form of reasoning called "hypothetico-deduction", that can be used to address the problem of multiple explanations which arises in the application of abduction to knowledge assimilation and diagnosis.

In a framework of hypothetico-deductive reasoning the knowledge is split into the theory T and observable relations S which may be tested through experiments. The basic idea behind the reasoning process is to formulate and decide between alternative hypotheses. This is performed through an interaction between the theory and the actual observations. The technique allows this interaction to be user mediated, permitting the acquisition of further information through experimental tests. Abductive explanations which have all their empirical consequences observed are said to be "fully corroborated".

We set up the basic theoretical framework for hypothetico-deductive reasoning and develop a corresponding proof procedure. We demonstrate how hypothetico-deductive reasoning deals with one of the main characteristics of common-sense reasoning, namely incomplete information, through the use of partial corroboration. We study the extension of basic hypothetico-deductive reasoning applied to theories that incorporate default reasoning as captured by negation-as failure (NAF) in Logic Programming. This is applied to the domain of Temporal Reasoning, where NAF is used to formulate default persistence. We show how it can be used successfully to tackle typical problems in this domain.

1 Motivation

Abduction is commonly adopted as an approach to diagnostic reasoning [Reggia & Nau, 1984], [Poole, 1988]. However, there are frequently many possible abductive explanations for a given observation. This is the problem of "multiple explanations". In order to choose between these explanations it becomes necessary to collect more information. Consider the Crime Detection example formalized below (Theory T1).

Suppose we arrive at the scene of the crime and the first observation we make is that someone is dead. We seek an explanation for this on the basis of the theory T1 above. Suppose we accept that there are only three possible causes of death: being strangled, being stabbed, or drinking arsenic (these are technically known as the *abducibles*). Simple abduction starting from the observation "dead" yields precisely these three possible explanations. In order to choose between these multiple explanations, we need to collect more information. For example, if we examined the corpse and discovered that there were marks on the neck, we might take this as evidence for the first explanation over the others. Moreover, we know that drinking arsenic also has the consequence of leaving the victim with a blue tongue, so we might like to look for that.

Theory T1

strangled → dead	strangled → neck_marks
blood_loss → dead	stabbed → blood_loss
poisoned → dead	drunk_arsenic → poisoned
drunk_arsenic → blue_tongue	

One approach to deciding between multiple explanations is through the performance of *crucial experiments* ([Sattar & Goebel, 1989]): pairs of explanations are examined for contradictory consequences, and an experiment is performed which *refutes* one of them whilst simultaneously *corroborating* the other. With *n* competing explanations we must thus perform at most *(n-1)* crucial experiments .

The crucial experiment approach is, however, unable to choose between explanations when they fail to have contradictory consequences or when they have contradictory consequences that are not empirically determinable (e.g. Tychonic and Copernican world systems). In our example, for instance, the explanations "strangled" and "stabbed" are not incompatible. It is possible that the victim was *both* strangled *and* stabbed. As result, there can be no crucial experiment that will decide between the two. However, further evidence might lead us to accept one explanation, whilst tentatively rejecting the other. For example, knowledge that the person exhibits marks on the neck supports the "strangled" hypothesis. In fact we have all the theoretically necessary observations to conclude that the victim was strangled. On the other hand, the "stabbed" hypothesis implies "blood_loss", which if not observed might lead us to favour the "strangled" explanation. Note that later evidence of blood loss would lead us to return to the "stabbed" hypothesis (in addition to "strangled"). From our viewpoint, crucial experiments are the special case of general hypothetico-deductive reasoning when an hypothesis is refuted whilst simultaneously corroborating a second.

The process of hypothetico-deductive reasoning allows the formation and testing of hypotheses within an interactive framework which is applicable to a wide class of applications and is implementable using existing technology for resolution.

The technique of hypothetico-deductive reasoning has its origin in the Philosophy of Science. It was primarily proposed by opponents of Scientific Induction. Its notable contributors were Karl Popper ([Popper, 1959],[Popper, 1965]), and Carl Hempel [Hempel, 1965]. In its original context, hypothetico-deduction is a method of creating scientific theories by making an hypothesis from which results already obtained could have been deduced and which entails new predictions that can be corroborated or refuted. It is based on the idea that hypotheses cannot be derived from observation, but once formulated can be tested against observation.

The hypothetico-deductive mechanism we formulate, resembles this method in having the two components of hypothesis formation and corroboration. It differs from the accepted usage of the term in philosophy of science by the status of the hypothesis formation component.

In the philosophy of the process of hypothesis formation is equivalent to theory formation: a creative process in which a complete theory is constructed to account for the known observations. By contrast, the method we describe here starts with a fixed generalized theory which is assumed to be complete and correct. The task is to construct some hypotheses which when added to the theory have the known observations as logical consequences. The process is more akin to that used by an engineer when they apply classical mechanics to a particular situation: they don't seek a new physical theory, but rather a set of hypotheses which would explain what they have observed. Since, for us, hypothesis formation can be mechanized, we do not have to tackle the traditional issues of the philosophy of science concerning the basis of theory formation. We thus avoid (like Poole before us [Poole, 1988, p.28]) one of the most difficult problems of science.

This paper is organized as follows. We first describe the reasoning process and present the logical structure of the reasoning mechanism, indicating how it relates to classical deduction and model theory. Abductive and corroborative derivation procedures for implementing the reasoning process are then defined through resolution. We indicate how this reasoning technique relates to current work on abduction and diagnostic reasoning, and suggest some possible extensions. We illustrate the features and applicability of this reasoning method with several examples. We then describe the extension of hypothetico-deduction to apply to theories which include some form of default reasoning, using negation-as-failure as an example. We consider a typical application of defaults in causal reasoning, namely default persistence, and provide several further examples which illustrate this extension.

2 Hypothetico-deductive Framework

Suppose we have a fixed logical theory T about the world. For example, it might be a medical model of the anatomy, or a representation of the connections in an electrical network, or a model of the flow of urban traffic in Madrid. Let us divide the relations in the theory into two categories: empirical and theoretical. How we make this distinction will depend on how we interpret these relations in the domain for the theory. An empirical relation is one which can be (or has been) observed. For example, the blood pressure of a patient, the status of a circuit-breaker (open or closed), or the number of cars passing some point. By contrast, a theoretical relation is in principle not observable. Examples of theoretical relations might be infection with an influenza virus, the occurrence of a short-circuit from the viewpoint of a control centre, or the density of traffic at some point.

Suppose we want an explanation for G on the basis of the theory. By this, what we mean is "what relations (we will call them *hypotheses*) might be true in order to have given rise to G?". The answer to this question could involve either theoretical or empirical relations. In order to be confident that an explanation is the *correct* explanation it is useful to *test it*. Explanations in terms of empirical relations are directly testable. In the simplest case we just consider the other observations we have already made; in more complicated cases, we may need to "go and look" or even perform an "experiment". Explanations in terms of theoretical relations must be tested indirectly, by deducing their empirical consequences, and testing these.

Unfortunately, not all hypotheses that might give rise to the observation G serve as explanations, regardless as to whether they pass any tests. Some are too trivial such as taking G as an explanation for itself. Others we rule out as unsuitably shallow. For example, suppose we sought an explanation for the observation "Jo laughed at the joke"; one possible hypothesis is because "the joke was funny". However, what we really wanted was a deeper explanation: Why was the joke funny? We therefore designate certain types of hypotheses as explanatory (or, more strictly, "abducible").

The problem of explanation, as far as we are concerned in this paper, is the problem of constructing abducible hypotheses which when we add them to T will have G as a logical consequence. Furthermore, explanations must pass (direct or indirect) tests.

The process of constructing hypotheses which have G as a deductive consequence is an example of **hypothesis formation**. It is this stage that corresponds to the "hypothetico-" component of hypothetico-deductive reasoning. The process of testing an explanation is an example of **corroboration**. It is this stage that corresponds to the "deductive" component of hypothetico-deductive reasoning. This is because we use deduction to determine the empirical consequences of a given explanation. The process of hypothetico-deductive reasoning can now be formulated as the construction of an explanation for an observation through interleaving hypothesis formation and corroboration.

3 The Hypothetico-deductive Mechanism

Let us consider the mechanism for hypothetico-deductive reasoning in more detail. To simplify matters we shall require that our theory is composed of rules and no facts. In logical terms, an hypothesis (and thus an explanation) will be a set of ground atomic well-formed formulae.

Suppose we have a (usually causal) theory T, an observation set O, a set of abducible atomic formulae A, and a particular observation G from O which we wish to explain. Let O′ = O-G. In addition we define a set S, the **observables**, containing all the formulae that can occur in O.

There are three components to the reasoning process: hypothesis formation, hypothesis corroboration, and explanation corroboration. In outline, we carry out hypothesis formation on G, and for each component formula in the resultant hypothesis. We repeat this process until all that remains is a set of abducible relations constituting the explanation. We also carry out hypothesis corroboration at each formation point. Finally we reason forwards from the explanation to perform explanation corroboration.

Hypothesis Formation

From any ground atomic formula F we form an hypothesis for that formula. This is done by determining which rules in T might allow F as a conclusion, and forming an hypothesis from the antecedents of each such rule (after carrying out the relevant substitutions dictated by F). Each hypothesis is thus sufficient to allow the conclusion of F.

Hypothesis Corroboration

An hypothesis for an observation may contain instances of observables defined by S. For each such component we check to see whether it is an observation recorded in O′. If it is a member of O′ then it is corroborated and we can retain it. However, where any component is not corroborated in this fashion, we reject the entire hypothesis.

Explanation Corroboration

An hypothesis H which is composed entirely of instances of abducible predicates defined by A is an **explanatory hypothesis.** To corroborate H, we use T to reason forwards from H as an assumption. Each logical consequence of H which is also an instance of an observable is checked against O′ for corroboration (similar to "hypothesis corroboration"). If it does not occur in O′ then the original hypothesis H is rejected. If all observable consequences are corroborated, then the explanation H is said to be corroborated.

In general, rules may have more than one literal in their antecedent. We must also check the satisfaction of the other literals in a given rule by reasoning backwards until we reach either one of the observations in O′ or one of the other explanatory hypotheses. If neither of these two situations arise, the rule is discarded from the forward reasoning process.

We make a distinction between corroboration *failure*, where an hypothesis or prediction does not occur in the observation set O′, and *refutation*, where the negation of an hypothesis or prediction occurs in O′. Normally the form of O and T means that refutation is impossible (see the next section for details of this form). Later we suggest an extension which allows the possibility of refutation in addition to corroboration failure. In cases where it is natural to apply the closed world assumption to O, these two situations will coincide.

4 The Logical Structure of Hypothetico-deductive Reasoning

Suppose we have a theory T composed of definite Horn clauses and an observation set of ground atomic well-formed formulae O. Let the set of ground atomic formulae which can occur in O be S, the **observables.** Similarly, let us define a set of distinguished ground atomic formulae A, the **abducibles,** in terms of which all explanations must be constructed. An explanation will be a member of the set A. We will assume that the theory T alone does not entail any empirical observation without some other empirical input i.e. there does not exist any formula ϕ such that $\phi \in S$ and $T \models \phi$. Consider also a ground atomic formula G (a member of S) for which we seek an explanation.

Given the 4-tuple <T,O,A,S>, a **corroborated explanation** Δ for G, is a set of ground atomic well-formed formulae, which fulfils all of the following criteria:

(1) Each formula in Δ must be a member of A.

(2) $T \cup \Delta \models G$

(3) If $T \cup \Delta \models \Pi$ and $\Pi \subseteq S$, then $\Pi \subseteq O$

An explanation set Δ which satisfies (1) and (2) but not (3) is said to be **uncorroborated.**

This formulation is easily generalized to explanation for multiple observations by simply replacing G with a conjunction of ground atomic formulae.

We note that since at this stage we have taken our theories to be Horn, a simple extension to hypothetico-deductive reasoning allows us to distinguish between explanation *refutation* when a prediction is inconsistent with observation, and merely the failure of *corroboration* where a prediction is consistent with known observations but not present in them. Such an extension would allow a hypothetico-deductive system to deal with circumstances where our observations cannot ever be complete (where we know our fault-detection system is itself fallible, for instance). We could then discard only those explanations that are refuted, and order the remaining ones according to their degree of corroboration (corresponding to Popper's notion of *versimilitude*, [Popper, 1965]). A later section discusses the extension of hypothetico-deductive reasoning to theories which include negation-as-failure.

This extended version of hypothetico-deductive reasoning is non-monotonic because later information might serve to refute a partially corroborated explanation. To return to our first example for instance, the observation that the victim does not have a blue tongue would lead us to reject the hypothesis that they had drunk arsenic (even if previously this hypothesis had some observational consequences which had been observed).

5 Hypothetico-deductive Proof Procedure

A resolution proof procedure which implements hypothetico-deductive reasoning is formally presented below. Basically we define two types of derivation: abductive derivation and corroboration derivation which are then interleaved to define the proof procedure. Abductive derivation corresponds to the processes of hypothesis formation and corroboration, deriving hypotheses *for* goals. Corroboration derivation corresponds to the process of explanation corroboration, deriving predictions *from* goals. There are two different ways to interleave the abductive and deductive components of the reasoning mechanism. One approach is to derive all the abducible literals in the hypothesis for an observation, *before* any of them are corroborated. The second approach attempts corroboration as soon as an abducible literal is derived, postponing consideration of other (non-abducible) literals in the hypothesis. Here we present a proof procedure based on the second approach.

Definition (*safe selection rule*)

A safe selection rule R is a (partial) function which, given a goal $\leftarrow L_1, \ldots, L_k$ $k \geq 1$ returns an atom L_i, $i=1,\ldots,k$ such that:

either i) L_i is not abducible;

or ii) L_i is ground.

Definition (*Hypothetico-deductive proof procedure*)

An ***abductive derivation*** from $(G_1\ \Delta_1)$ to $(G_n\ \Delta_n)$ via a safe selection rule R is a sequence

$$(G_1\ \Delta_1),\ (G_2\ \Delta_2),\ \dots\ ,(G_n\ \Delta_n)$$

such that for each $i>1$ G_i has the form $\leftarrow L_1,\dots,L_k$, $R(G_i)=L_j$ and $(G_{i+1}\ \Delta_{i+1})$ is obtained according to one of the following rules:

A1) If L_j is neither an abducible nor an observable, then $G_{i+1}=C$ and $\Delta_{i+1}=\Delta_i$ where C is the resolvent of some clause in T with G_i on the selected literal L_j;

A2) If L_j is observable, then $G_{i+1}=C$ and $\Delta_{i+1}=\Delta_i$ where C is the resolvent of $C' : \leftarrow L_1',\dots,L_j',\dots,L_k'$ with some clause in T on L_j' where $\leftarrow L_1',\dots,L_{j-1}',L_{j+1}',\dots,L_k'$ is the resolvent of G_i with some clause (ground assertion) L_j' in O on the selected literal L_j;

A3) If L_j is abducible and $L_j\in \Delta_i$, then $G_{i+1}= \leftarrow L_1,\dots,L_{j-1},L_{j+1},\dots,L_k$ and $\Delta_{i+1}=\Delta_i$;

A4) If L_j is abducible and $L_j\notin \Delta$ and there exists a **corroboration derivation** from $(\{L_j\}\ \Delta_i\cup\{L_j\})$ to $(\{\}\ \Delta')$ then $G_{i+1}= \leftarrow L_1,\dots,L_{j-1}, L_{j+1},\dots,L_k$ and $\Delta_{i+1} = \Delta'$.

Step A1) is an SLD-resolution step with the rules of T. ~~In step A2) under the assumption that~~ observables and abducibles are disjoint we need to reason backward from the true observables in the goal to find explanations for them since the definition of an explanation requires that it logically implies G in the theory T alone without the set of observations O. Step A3) handles the case where an abductive hypotheses is required more than once. In step A4) a new abductive hypotheses is required which is added to the current set of hypotheses provided it is corroborated.

A ***corroboration derivation*** from $(F_1\ \Delta_1)$ to $(F_n\ \Delta_n)$ is a sequence

$$(F_1\ \Delta_1),\ (F_2\ \Delta_2)\ \dots\ (F_n\ \Delta_n) \text{ to } (F_n\ \Delta_n)$$

such that for each $i>1$ F_i has the form $\{H\leftarrow L_1,\dots,L_k\} \cup F_i'$ and $(F_{i+1}\ \Delta_{i+1})$ is obtained according to one of the following rules:

C1) If H is not observable then $F_{i+1} = C' \cup F_i'$ where C' is the set of all resolvents of clauses in T with $H\leftarrow L_1,\dots,L_k$ on the atom H and $\Delta_{i+1}=\Delta_i$;

C2) If H is a ground observable, $H\notin O$ and $L_1,\dots,L_k$ is not empty then $F_{i+1} = C' \cup F_i'$ where C' is $\leftarrow L_1,\dots,L_k$ and $\Delta_{i+1}=\Delta_i$; If $H\in O$ then $F_{i+1} = F_i'$ and $\Delta_{i+1}=\Delta_i$.

C3) If H is a non ground observable, $O\nvDash \exists x H$ and $L_1,\dots,L_k$ is not empty then $F_{i+1} = C' \cup F_i'$ where C' is $\leftarrow L_1,\dots,L_k$ and $\Delta_{i+1}=\Delta_i$;

C4) If H is a non ground observable and L_j is any non observable selected literal from $L_1,\dots,L_k$ then $F_{i+1} = C' \cup F_i'$ where C' is the set of all resolvents of clauses in $T \cup \Delta_i$ with $H\leftarrow L_1,\dots,L_k$ on the selected literal L_j and $\Delta_{i+1}=\Delta_i$; If L_j is observable the resolutions are done only with clauses in O.

C5) If H is empty, L_j is any selected literal and L_j is not observable then $F_{i+1} = C' \cup F_i'$ where C' is the set of all resolvents of clauses in $T \cup \Delta_i$ with $\leftarrow L_1,\dots,L_k$ on the literal L_j and $\Box\notin C'$, and $\Delta_{i+1}=\Delta_i$; If L_j is observable the resolutions are done only with clauses in O.

In step C1) we "reason forward" from the conclusion H trying to generate a ground observable at the head. Once this happens if this observable is not "true" steps C2), C3) give the denial of the conditions that imply this observable. Step C4) reasons backward from the conditions either failing or trying to instantiate further the observable head. Step C5) reasons backward from the denials of steps C2), C3) until every possible such backward reasoning branch fails. Note that in the backward reasoning steps observables are resolved from the observations O and not the theory. More importantly notice that we do not reason forward from an observable that is true.

Note that we have included the set of hypotheses Δ_i in the definition of the corroboration derivation although this does not get affected by this part of the procedure. The reason for this is that more efficient extensions of the procedure can be defined by adding extra abducible information in the Δ_i during the corroboration phase e.g.the required absence of some abducible A can be recorded by the addition of a new abducible A^*.

Theorem

Let $<T,O,A,S>$ be a Hypothetico-Deductive framework and G a ground atomic formula. If $(\leftarrow G\ \{\})$ has an adbuctive derivation to $(\Box, \Delta)$ then the set Δ is a corroborated explanation for G.

Proof (Sketch)

The soundness of the abductive derivations follows directly from the soundness of SLD resolution for definite Horn theories as every abductive derivation step of this procedure can be mapped into an SLD resolution step. To show that the explanation Δ is corroborated let $A\in S$ be any ground atomic logical consequence of $T \cup \Delta$. Since $T \cup \Delta$ is a definite Horn theory A must belong to its minimal model which can be constructed in terms of the immediate consequence operator $\mathcal{T}$ [van Emden & Kowalski, 1976] . Hence there exists a finite integer n such that $A \in \mathcal{T}_{T\cup\Delta}\uparrow^{n}(\varnothing)$ and A does not follow from T alone by our assumption on the form of the theory T . The result then follows by induction on the length of the corroboration derivation.

6 Application of Hypothetico-deductive Reasoning

In this section we will illustrate hypothetico-deductive reasoning with some examples. Before this it is worth pointing out that existing abductive diagnosis techniques (e.g. [Poole et al., 1987], [Davis, 1984], [Cox & Pietrzkowski, 1987], [Genesereth, 1984], [Reggia et al., 1983], [Sattar & Goebel, 1989]) can be accommodated within the HD framework. For example in the diagnosis of faults in electrical circuits hypothetico-deductive reasoning exhibits similar behaviour to [Genesereth, 1984], [Sattar & Goebel, 1989].

Problems and domains which are ideally suited to the application of hypothetico-deductive reasoning exhibit two characteristics. Firstly, they have a large number of possible explanations in comparison to the number of empirical consequences of each of those explanations. Secondly, they have a minimal amount of observational data pertaining to a given explanation so that corroboration failure is maximized.

To illustrate the manner in which general hypothetico-deductive reasoning deals with differing but compatible explanations, let us consider the example of abdominal pain first presented by [Pople, 1985] and axiomatized in [Sattar & Goebel, 1990]. The axioms are reproduced below. To allow the possibility of several diseases occurring simultaneously, the three expressions which capture the fact that the symptoms (nausea, irritation_in_bowel, and heartburn) are incompatible, have been omitted.

Theory T2

abdominal_pain_symp(X) → has_abdominal_pain
problem_is(indigestion) → abdominal_pain_symp(nausea)
problem_is(dysentry) →
abdominal_pain_symp(irritation_in_bowel)
problem_is(acidity) → abdominal_pain_symp(heartburn)

Now consider the following observations:

Observations O

has_abdominal_pain
abdominal_pain_symp(nausea)

Abducibles, **A** = {problem_is(indigestion),
problem_is(dysentry),
problem_is(acidity)}

Observables, **S** =
{has_abdominal_pain,
abdominal_pain_symp(nausea),
abdominal_pain_symp(irritation_in_bowel),
abdominal_pain_symp(heartburn)}

There are three possible potential explanations for the observation "has_abdominal_pain". Since they are not mutually incompatible (it is possible to have all three diseases, for example), there is no crucial literal which can help us distinguish between the three explanations. There is thus no "best" explanation from this point of view.

From the point of view of hypothetico-deductive reasoning however, one of the explanations stands apart from the others. On the basis of all the currently available evidence "problem_is(indigestion)" is completely corroborated. The two remaining explanations remain possible but uncorroborated; that is to say there is no supplementary evidence in support of them. Experiments might be performed (testing for "abdominal_pain_symp(irritation_in_bowel)", and "abdominal_pain_symp(heartburn)") which could corroborate one or both of the others, which would lead us to extend our explanation. Since physical incompatibilities are rare in common-sense reasoning, hypothetico-deductive reasoning has an advantage in being able to offer a (revisable) "best" explanation based on the currently available evidence, in spite of the absence of possible crucial experiments. It is important to appreciate that it is usually impractical to simply construct the hypotheses by performing abduction on all the observations in O, since in general there may be an extremely large number of them. Moreover, only a few may be relevant to the particular observation for which we seek an explanation.

It might be thought that the checking of *all* the observational consequences of some explanation might be equally impractical: there might be an infinite number of them as well. However, it must be borne in mind that we are only considering the representation of common-sense; we would normally ensure that there are only a small number of observable consequences in which we would be interested. We would define our set of observables, *S*, accordingly. So, for instance, in the fermentation example below we represent certain critical times (often referred to as "landmarks") at which we might perform observations. Similarly, in the "stolen car" example which we present later, we restrict observables to events that occurred at some specific point in time.

One application area in which incomplete information is intrinsic, is that of temporal reasoning. Reasoning about time is constrained by the fact that factual information is only available concerning the past and the present. By its very nature we must perform temporal diagnosis with no knowledge about the future states of the systems we are trying to model.

As an example of temporal diagnosis which illustrates this characteristic, consider an industrial process involving the fermentation of wine. Suppose we are faced with the task of diagnosing whether the fermentation process has proceeded normally, or that the extremely rare conditions have occurred under which we will produce a vintage wine. To do this we must carry out a test at some time after the wine-making process has begun, such as measuring its pH, its relative density, or its alcohol content. Suppose further that we need to decide on this diagnosis before a certain time, e.g. the bottling-time tomorrow. Let us refer to some property of the mixture which would be observed for vintage wine by the symbol **p1**, and that for ordinary wine as **p2**. These two properties might be entirely compatible: it is perfectly possible for ordinary wine to be produced under conditions which exhibit **p1**(as well as **p2**), but in such a case it is not the fact that the mixture is ordinary wine that *causes* **p1** to be observed. Now suppose we observe **p1** before the bottling time, and suppose there are no further observational consequences for the "vintage wine" hypothesis that are observable before tomorrow. Then the "vintage wine" hypothesis is completely corroborated within the defined time-scale. On the other hand, the "ordinary wine" hypothesis remains at best only partially corroborated. Hypothetico-deductive reasoning would then *prefer* the "vintage wine" hypothesis over the "ordinary wine" one. The temporal dimension illustrates the ability of hypothetico-deductive reasoning to form diagnoses on the basis of incomplete information. Notice that an extension of the time scale would revise the status of the observable relations and perhaps the "vintage wine" hypothesis would become only partially corroborated. The application of hypothetico-deductive reasoning to the temporal domain will be discussed in more detail in the next section as an important special case of the integration of hypothetico-deductive reasoning and default reasoning.

7 Hypothetico-deduction with Default Theories

As we discussed above, the aim of hypothetico-deductive reasoning has been to provide a framework in which we can tackle one of the main characteristics of common sense reasoning, namely incomplete information. More specifically it addresses the fact that we are often forced to form hypotheses and explanations on the basis of limited information. Another important form of reasoning that deals with the problem of incomplete (or limited) information is default reasoning (see e.g. [Reiter, 1980]). We can then enhance the capability of each framework separately to deal with this problem of missing information by integrating them together into a common framework.

So far we have only considered the application of hypothetico-deduction to classical theories. In this section we study its application to default theories incorporating negation-as-failure (NAF) from Logic Programming. We will then apply this adaptation of hypothetico-deduction to temporal reasoning problems formulated within the event calculus where NAF is used to represent default persistence in time ([Kowalski & Sergot, 1987], [Evans, 1989]).

The approach we adopt is to consider only classical theories to which non-monotonic reasoning mechanisms such as default and hypothetico-deductive reasoning are applied (in contrast to non-monotonic logics). The motivation as before, is to separate representation (classical logic) from reasoning (non-monotonic). Recent formalizations of the semantics of negation-as-failure [Eshghi & Kowalski, 1989], [Kakas & Mancarella, 1990], [Dung, 1991], [Kakas & Mancarella, 1991] have adopted a similar point of view. This approach means that hypothetico-deductive reasoning can be applied to default theories of any system which separates these two components, e.g. circumscription [McCarthy, 1980].

Following this work, we associate to any general logic program, P, (Horn clauses extended with negation-as-failure) a classical theory , P′, as follows. Each negative condition, **not** p, where **not** denotes the negation-as-failure operator, is regarded as a single new positive atom. This can be made explicit by replacing each such negative literal, **not** p, by a syntactic variant, say p*, to give the Horn theory P′. The model-theoretic extension of the new symbol is intended to be the complement of the old one, so that we can omit the **not**. To take a more meaningful example we might replace "**not** alive" with "dead". These new symbols "p*" or "dead" are then defined to be abducible predicates. The above authors show that with this view it is possible to understand (and generalize) the stable model semantics [Gelfond & Lifschitz, 1989] for NAF in logic programming. (Note that this is also the approach taken more generally in [Poole, 1988] for understanding default reasoning through abduction by naming the defaults and considering these as assumptions.)

We can then apply an adapted formulation of hypothetico-deductive reasoning to these classical Horn theories P′ corresponding to general logic programs P. As above we have a 4-tuple <P′,O,A,S> where the set, A, of abducibles has been extended with new abducibles e.g. "p*", "dead", which name the different NAF default assumptions.

Hence given a 4-tuple <P′,O,A,S>, a **corroborated explanation** Δ for an observation G, is a set of ground atomic well-formed formulae, which fulfils all of the following criteria:

(1) Each formula in Δ is a member of A.
Let $\Delta = \Delta_D \cup \Delta_H$ where Δ_D denotes the subset of abducibles corresponding to NAF.

(2) $P' \cup \Delta \models G$

(3) If $P' \cup \Delta \models \Pi$ and $\Pi \subseteq S$, then $\Pi \subseteq O$

(4) There exists a stable model[1] M of $P' \cup \Delta_H \cup O$ such that the negations corresponding to Δ_D hold in M (i.e. are contained in the complement of M).

This is a direct extension of the previous definition of hypothetico-deductive reasoning. The extra condition (4) captures the default reasoning present in the theory P (or P′). This is clearly separated in this condition although it does play an important role in the generation of explanations by rejecting explanations that do not satisfy it. This has the effect of adding extra abducibles in the Δ to make it acceptable. For example in the theory,

$$\begin{array}{lcl} G & \leftarrow & p^* \\ p & \leftarrow & q^* \\ q & \leftarrow & a \end{array}$$

although {p*} is an explanation for G, this is not accepted until the abducible "a" is added to it which ensures that this default assumption {p*} is valid. In addition condition (4) also ensures that any default assumption (abducible) in Δ is compatible with the observations O. Note that we could have chosen to put together conditions (2) and (4) as "G is true in a stable model of $P \cup \Delta_H$" for generating the explanations Δ, and use condition (4) solely for the purpose of ensuring that Δ_D are compatible with the observations O.

Although at first sight it might seem appropriate to allow default reasoning during the corroboration of an explanation this is not the case as indicated by condition (3). The reason for this is clear: if we allow it then the corroboration process will not be for the explanation Δ alone, but for Δ plus any additional default assumptions made in arriving at the observable test. In other words, we would not want to reject an explanation Δ by failure to corroborate an observation that is a not a consequence of Δ alone but of Δ with some additional default assumptions.

Let us now indicate how the proof procedure for hypothetico-deductive reasoning, defined earlier, needs to be extended to deal with this more general formulation where our theories are general logic programs. The first thing to notice is that, as indicated by condition (3), the corroboration phase of the procedure remains unchanged apart from the fact that it will also be applied whenever a NAF hypothesis, "p*" (or "**not** p"), is added to the explanation. Similarly, the adbuctive derivation phase remains as before with the set of abducibles enlarged to include the NAF default assumptions.

The main extension of the procedure arises from the need to implement the new condition (4). This can be done by adopting the abductive proof procedure developed in [Eshghi & Kowalski, 1989], [Kakas & Mancarella, 1990b], [Kakas & Mancarella, 1990c] for NAF which is an extension of SLDNF. A new type of derivation, called **consistency derivation**, is introduced interleaved with the abductive phase of the procedure whenever a NAF hypothesis, "p*" (or "**not** p"), is required in the explanation. Its purpose is to ensure that "p*" (or "**not** p") is a valid NAF assumption by checking that p does not succeed. This involves reasoning backwards from p in all possible ways and showing that each such branch ends in failure.

During this consistency check for some NAF hypothesis, "p*" (or "**not** p"), it is possible for new

[1] More generally, we can use recent extensions of stable models e.g. preferred extensions or stable theories as defined in [Dung, 1991] and [Kakas & Mancarella, 1991] respectively.

abductive phases to be generated whenever the failure of some consistency branch reduces to showing that some other NAF default assumption e.g. "q*" (or "**not** q") does not hold in the theory $P' \cup \Delta$. To ensure this the procedure starts a new abductive phase to show that q holds where it is possible that new hypotheses may be added in the explanation if this is needed to prove q. Then with this enlarged explanation "q*" (or "**not** q") is not a valid (default) NAF assumption (as q holds) and so the original consistency branch can not succeed. In the example above the abducible "a" in the explanation {p*, a} for G is generated during the consistency check of p* (or **not** p) as described here. More details about this extension of the proof procedure can be found in the references above.

8 Application of HD Reasoning to Temporal Reasoning

As an example of the application of the above extended hypothetico-deductive mechanism, let us consider temporal reasoning with the Event Calculus [Kowalski & Sergot, 1987] where NAF is used to express default persistence in time.

The Event Calculus represents *properties* which hold over intervals of time. They are initiated and terminated by *events* which happen at particular instances of time. NAF is used to conclude that a property is not "clipped" or "broken" over an interval of time, achieving default persistence. Variants of the two main axioms, which define when a property "holds" and when a property is "broken", are given below.

$$\begin{aligned} \text{holds-at}(p,t2) \leftarrow\ & \text{happens-at}(e,t1) \wedge \\ & \text{initiates}(e,p) \wedge \\ & t1 < t2 \wedge \\ & \textbf{not}\ \text{broken-during}(p,<t1,t2>) \\ \text{broken-during}(p,<t1,t2>) \leftarrow\ & \text{happens-at}(e,t) \wedge \\ & \text{terminates}(e,p) \wedge \\ & t1 < t \wedge \\ & t \leq t2 \end{aligned}$$

The first axiom states that some property p holds at any time after an initiating event, provided it is not (known to be) broken at some time during the intervening time-interval. NAF ensures that we draw the conclusion that it isn't broken if we have no evidence for it: default persistence. The second axiom states that a property is broken during an interval if a terminating event happens at some time within that interval.

Before we can apply HD reasoning to these axioms we must carry out the transformation to eliminate the NAF. A possible renaming of "not broken-during" is "persists":

$$\begin{aligned} \text{holds-at}(p,t2) \leftarrow\ & \text{happens-at}(e,t1) \wedge \\ & \text{initiates}(e,p) \wedge \\ & t1 < t2 \wedge \\ & \text{persists}(p,<t1,t2>). \end{aligned}$$

Before we present a detailed example of the application of HD, let us briefly consider how the use of a temporal default theory such as the Event Calculus does not modify the process of corroboration (we use the classical version of the theory), although it does modify the process of explanation construction.

Consider an example in which the walls of a house are painted white. Using the Event Calculus, if we wished to explain why the walls were white, we would hypothesize an event of painting them white. In order to corroborate this hypothesis we would look for empirical consequences. One possibility might be that the paint brush has white paint on it. However this prediction involves assuming that the state of "brush-has-white-paint" persisted since the walls were painted; the corroboration is based upon a further (uncorroborated!) hypothesis. Moreover, consider the consequences of observing that the paint brush has *red* paint on it. Does this *refute* the explanation that the walls are white because they were painted white? Obviously not. Under the extended HD scheme we limit default reasoning to be a part of the hypothesis formation component. Corroboration is straightforward classical deduction. This is one of the reasons for having to transform the Event Calculus axioms to eliminate the NAF.

Let us consider a more detailed application of hypothetico-deductive reasoning to a problem formalized in the Event Calculus. We shall take Kautz's "stolen car" problem [Kautz, 1986]. The task is to explain why a car parked in the morning is missing when we look for it in the afternoon. In particular, to explain *when* the car was stolen. Kautz's original motivation was to demonstrate that temporal reasoning which performed chronological minimization (e.g. Shoham's Non-monotonic Logic [Shoham, 1988]) would predict that the car was stolen the instant before it was found to be missing; which was unsatisfactory. From our point of view, the stolen car problem is more correctly viewed as an *explanation* problem in which there are several possible competing explanations, corresponding to the different times that the car might have been stolen.

In the formalism of the Event Calculus we would describe the problem as follows. We know that the car was parked at some particular time, say time "1"; and we know that it was missing at, say, time "4". We also know that stealing initiates the property "missing" and terminates "parked":

$$\begin{aligned} \text{initiates}(e,\text{missing}) &\leftarrow \text{type}(e,\text{steal}) \\ \text{terminates}(e,\text{parked}) &\leftarrow \text{type}(e,\text{steal}) \end{aligned}$$

Our explanatory task is thus to explain the observation "holds-at(missing,4)". We will take the predicates "happens-at", "type" and (since it is a default relation) "persists" to be abducible. Furthermore, let us restrict the abducible "happens-at" events to those which happen between time "1" and "4". Our observables will be instances of the relation "holds-at" which occur at time "4". Using hypothesis formation applied to the rule defining "holds-at" we might hypothesize:

$$\{\text{happens-at}(e,2),\ \text{type}(e,\text{steal}),\ \text{persists}(\text{missing},<2,4>)\}$$

This states that some stealing event happened at time "2". Notice that we have to include the persistence assumption: if some other event had terminated this "missing" state (such as the returning of the car!), then this particular stealing event would not be the right explanation.

Using a discrete representation of time, there is another explanation corresponding to a stealing event at time "3". Pure abduction is unable to distinguish between these two explanations.

There are two further characteristics of HD to demonstrate. Firstly, note that we have to check the consistency of the default "persists" hypothesis (according to the 4th corroboration requirement). We do this by checking that "~broken-during(missing,<2,4>)" holds in the stable model when we include all our observations; computationally speaking, we must check that "broken-during(missing,<2,4>)" finitely fails.

The second characteristic is corroboration to choose between the two competing explanations. In order to describe this aspect, we must elaborate our example somewhat. Suppose that we had a car alarm fitted and it is not possible to steal the car without setting off the alarm. The hypothesis that the car was stolen at time "2" would lead us to predict "happens-at(alarm,2)" whereas the alternative would predict "happens-at(alarm,3)". We must extend our definition of observables to include "happens-at(alarm, 2)" and "happens-at(alarm, 3)", corresponding, say, to checking with someone near at what time they heard a car alarm start going off. The process of corroboration against observations concerning the alarm events proceeds as in the unextended version of HD reasoning.

Thus the addition of the appropriate observations for the "stolen car" situation allows us to form two explanations, one of which we might reject as uncorroborated and the other of which might be completely corroborated.

The "bloodless" Yale Shooting problem ([Morgenstern & Stein, 1988]) - the explanatory counterpart to the original Yale Shooting prediction problem ([Hanks & McDermott, 1987]) - is of a similar form. In this scenario, a gun is loaded, a period of waiting ensues, and someone is shot with the gun. They are found to be unharmed. The task is to explain how this could be so. Pure abduction produces a number of explanations in terms of unloading events that must have occurred during the period of waiting: one explanation for each different possible time of the event. Hypothetico-deduction allows the possibility of selecting one of the events as preferable on the grounds that it has empirical consequences which were observed.

9 Related and Further Work

Several authors have developed deductive techniques for the generation of hypotheses. In [Cox & Pietrzykowski, 1987] hypotheses are constructed from the terminal nodes of linear resolution proofs. Similarly, [Finger & Genesereth, 1985] perform "deductive synthesis" to provide "solutions to design problems" by "finding a *residue* for a given design goal"; and [Poole et al., 1987] use linear resolution for hypothesis generation implemented in the program THEORIST.

In [Eshghi & Kowalski, 1989], [Kakas & Mancarella, 1990] Horn clause logic programming is extended to include abduction with integrity constraints. The approach taken here, differs by the absence of integrity constraints although the process of checking abductive hypotheses by regarding them as updates, and reasoning forwards to integrity constraints, parallels the process of explanation corroboration we describe. There are two important differences between the application (rather than the technique) of explanation corroboration, and the integrity checking process. Firstly, we reason forwards to observables rather than integrity constraints; and secondly, the set of observables can be "dynamic". That is, we may have not made all the relevant observations: it may be necessary to perform an experiment to determine the outcome of corroboration (e.g. through "Query-the-user" [Sergot, 1983] in the case of an expert system). This approach of interactive acquisition of extra information to help decide between different explanations has been studied in [Kunifuji et al, 1986] in the context of Knowledge Assimilation. However, in some domainis of application it may be appropriate to use integrity constraints first for reducing the number of possible explanations before beginning the corroboration of explanations. The mechanisms developed in these papers are directly applicable to the incorporation of integrity checking in the hypothetico-deductive proof procedure defined above.

[Sattar & Goebel, 1989] describes how the THEORIST system can be extended through the notion of performing crucial experiments [Popper, 1965] using "crucial literals" (from [Seki & Takeuchi, 1985]) to decide between competing explanatory hypotheses. As we have mentioned above, this can be understood as special case of explanation corroboration used to decide between multiple incompatible explanations. The relative cost of carrying out the experimental tests for corroborating an explanation over the significance of this particular explanation is another feature that needs to be taken into account when further developing the hypothetico-deductive mechanism. For example, in circuit diagnosis [Davis, 1984] the failures are layered into categories according to their likelihood. De Kleer and Williams in [de Kleer & Williams, 1987] use probability and information theory to propose the next "best" test for localizing the fault in the framework of model based diagnosis. These techniques can be used to make our corroboration more efficient.

Conclusions

We have developed a versatile reasoning mechanism and proof procedure, based on the notion of corroboration, that is applicable to a variety of problems and logic-based systems in artificial intelligence. It combines the explanatory capability of hypothesis formation with the benefits of corroboration through deduction for control and testing. Hypothetico-deductive reasoning tackles the problem of undesired multiple explanations for an observation. It extends the isolated application of deductive and abductive reasoning. We have shown how the basic idea behind the reasoning process is to formulate and decide between alternative hypotheses. This is performed through an interaction between the theory and the actual observations. A suitable proof procedure for the implementation of hypothetico-deduction was presented. We have suggested that this form of reasoning might benefit for the use of a "query-the-user" facility. We have demonstrated how hypothetico-deductive reasoning deals with one of the main characteristics of common-sense reasoning, namely incomplete information, through the use of partial corroboration. Finally we have shown how the semantics of hypothetico-deduction can be extended to deal with default theories, in particular temporal theories such as the Event Calculus which include default persistence through the use of negation-as-failure. We have demonstrated how this extension can be applied to deal with Kautz's "stolen car" problem, and the "bloodless" counterpart to the Yale Shooting Problem.

Acknowledgements

We wish to thank R.A. Kowalski for useful discussions on the subject. Thank you, also, to Kate Moorcock for her common-sense examples. This work has been partially carried out under a long term project at Imperial College sponsored by FUJITSU and partially under Esprit project 2409, (EQUATOR). The EQUATOR partners are Ferranti Computer Systems Ltd. (UK), ERIA (Spain), Imperial College (UK), LABEN (Italy), SYSECA (France), CENA (France), CISE (Italy), PTT (Switzerland), SWIFT (Belgium), UCL (UK), ERITEL (Spain) and EPFL (Switzerland).

References

[Cox & Pietrzykowski, 1986] Cox, P. and Pietrzykowski, T. "Causes for Events: Their Computation and Applications"; Proc. CADE-86, J.Siekmann (ed.), Springer-Verlag, Lecture Notes in Computer Science, 1986, pp.608-621.

[Cox & Pietrzykowski, 1987] Cox, P. and Pietrzykowski, T. "General Diagnosis by Abductive Inference"; Technical Report, CS8701, School of C.S., University of Nova Scotia, 1987

[Davis, 1984] Davis, R. "Diagnostic Reasoning Based on Structure and Behaviour"; AI vol. 24, pp. 347-410, 1984.

[Dung, 1991] Dung P. M., Negation as Hypothesis; An Abductive Foundation for Logic Programming, in *Proc. 8th ICLP*, Paris, 1991.

[van Emden & Kowalski, 1976] van Emden M.H. and Kowalski R.A., The Semantics of Predicate Logic as a Programming Language, *Journal of ACM* 23, a (1976), pp. 733-742.

[Eshghi & Kowalski, 1989] Eshghi, K. and Kowalski, R. "Abduction Compared with Negation by Failure"; Proc. 6th ICLP, 1989.

[Finger & Genesereth, 1985] Finger, J. and Genesereth, M. "RESIDUE: A Deductive Approach to design synthesis"; Technical Report no. STAN-CS-85-1035, Dept. of Computer Science, Stanford University, 1985.

[Gelfond & Lifschitz, 1988] Gelfond, M., and Lifschitz, V. The Stable Model Semantics for Logic Programming; *Proceedings of the Logic Programming Conference*, Seattle, 1988.

[Genesereth, 1984] Genesereth, M. "The Use of Design Descriptions in Automated Diagnosis"; AI vol. 24, pp. 411-436, 1984.

[Hanks & McDermott, 1987] Hanks, S. and McDermott, D., Nonmonotonic Logic and Temporal Projection, in *Artificial Intelligence*, vol. 33, pp.379-412, 1987.

[Hempel, 1965] Hempel, C. "Aspects of Scientific Explanation and Other Essays in the Philosophy of Science"; The Free Press, New York, 1965.

[Kakas & Mancarella, 1990] Kakas, A.C. and Mancarella, P. "Generalized Stable Models: a Semantics for Abduction" In Proc. ECAI-90, 1990.

[Kakas & Mancarella, 1990b] Kakas, A.C. and Mancarella, P. "Database Updates through Abduction" in *Proc.16th International Conference on Very Large Data Bases, VLDB '90* , Brisbane, 1990.

[Kakas & Mancarella, 1990c] Kakas, A.C. and Mancarella, P. "On the relation of Abduction and Truth Maintenance" in *Proc. of the 1st Pacific Rim International Conference on AI, PRICAI-90,* Nagoya, Japan 1990.

[Kakas & Mancarella, 1991] Kakas, A.C. and Mancarella, P. "Stable Theories for Logic Programs", to appear in Proc. of ISLP-91, San Diego, 1991.

[Kautz, 1986] Kautz, H., "The Logic of Persistence" in Proc. AAAI-86, pp. 401, 1986.

[Kowalski & Sergot, 1986] Kowalski, R.A. and Sergot, M., A Logic-Based Calculus of Events, New Generation Computing, vol 4, pp. 267, 1986.

[Kunifuji et al, 1986] Kunifuji, S., Tsurumaki, K. and Furukawa, K., "Considerations os a Hypothesis-based Reasoning System" Journal of Japanese Society for Artificial Intelligence vol. 1 no. 2, pp.228-237, 1986.

[de Kleer & Williams, 1987] de Kleer, J. and Williams, B.C. "Diagnosing Multiple Faults"; Artificial Intelligence vol 32, pp. 97-130, 1987.

[McCarthy, 1980] McCarthy, J., Circumscription: A Form of Non-monotonic Reasoning; in *Artificial Intelligence*, vol. 13, pp.27-39, 1980.

[Morgenstern & Stein, 1988] Morgenstern, L. and Stein, L., "Why Things Go Wrong: A Formal Theory of Causal Reasoning"; Proc. AAAI '88, p.518ff.

[Poole, 1988] Poole, D. "Representing Knowledge for Logic-Based Diagnosis" In Proc.of the FGCS, pp.1282-1290, 1988.

[Poole et al.,. 1987] 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*, by N.Cercone and G.McCalla (ed.s), Springer-Verlag, New York, 1987, pp.331-352.

[Pople, 1985] Pople, H. "Coming to Grips with the Multiple Diagnosis Problem"; in*The Logic of Discovery and Diagnosis in Medicine*, Schaffner, K. (ed.), University of California Press, 1985.

[Popper, 1959] Popper, K. "The Logic of Scientific Discovery"; Basic Books, New York, 1959.

[Popper, 1965] Popper, K. "Conjectures and Refutations: The Growth of Scientific Knowledge";Harper Torch,New York, 1965.

[Reggia at al., 1983] Reggia, J., Nau, D., and Wang, P. "Diagnostic Expert System Based on a Set Covering Model"; Int. J. Man-machine Studies, vol. 19, pp.437-460, 1983.

[Reggia & Nau, 1984] Reggia, J.A. and Nau, D.S. " An Abductive Non-Monotonic Logic"; in Workshop on Non-Monotonic Reasoning, New Paltz, N.Y., 1984.

[Reiter, 1980] Reiter, R., "A Logic for Default Reasoning" Artificial Intelligence vol. 13, pp. 81-132, 1980.

[Sattar & Goebel, 1989] Sattar, A. and Goebel, R. "Using Crucial Literals to Select Better Theories"; Technical Report, Dept. of CS, University of Alberta, Canada, June 1989.

[Seki & Takeuchi, 1985] Seki, H. and Takeuchi, A. "An Algorithm for Finding a Query which Discriminates Competing Hypotheses"; Technical Report TR-143, Institute for New Generation Computer Technology, Tokyo, Japan, October 1985.

[Sergot, 1983] Sergot, M. "A Query-the-user Facility for Logic Programming"; Integrated Interactive Computer Systems, by P.Degano and E.Sandewell (ed.s), North Holland Press, pp.27-41.

[Shoham, 1988] Shoham, Y., "Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence"; MIT Press, 1988.