On Tests for Hypothetical Reasoning

Sheila McIlraith * and **Raymond Reiter**†
Department of Computer Science
University of Toronto
Toronto, M5S 1A4 Canada
email: mcilrait@ai.toronto.edu, reiter@ai.toronto.edu

Abstract

Suppose that HYP is a set of hypotheses which we currently entertain about some state of affairs represented by a propositional sentence Σ. In a diagnostic setting, HYP might consist of all the diagnoses of some device whose description is given by Σ, although our analysis is not restricted to diagnosis. Our concern is with tests – how they can be designed, and what conclusions can be drawn about the hypotheses in HYP as a result of performing tests. Specifically, we define the concept of a test and the concept of the outcome of a test. We characterize those tests whose outcomes refute or confirm an hypothesis, and discriminate between competing hypotheses. These characterizations are in terms of the prime implicates of Σ, and hence are implementable using assumption-based truth maintenance systems. In addition, we characterize the impact of a test outcome on consistency-based and abductive hypothesis spaces. Finally, we provide a characterization of differential diagnosis for consistency-based and abductive reasoning.

Introduction

In the AI literature on hypothetical reasoning there are relatively few results on the design of tests for discriminating a space of hypotheses, or on the conclusions one may draw from the outcome of a test. There are exceptions of course, particularly in the area of diagnosis. Among these are de Kleer and Williams [de Kleer and Williams, 1987] who provide a probabilistic analysis to decide what measurement to take next. The DART system of Genesereth [Genesereth, 1984] was capable of proposing circuit inputs and observations to be made in order to confirm or refute a possible diagnosis. TraumAID [Webber *et al.*, 1990] is a system for treating trauma patients which does sophisticated planning to design diagnostic tests and treatment. But by and large, there has been no systematic study of the design and role of tests in hypothetical reasoning. This paper is a first step in this direction.

Our concern in this paper is how tests provide information about the current space of hypotheses. Specifically, we define the concept of a test and the concept of the outcome of a test. We characterize those tests whose outcomes refute or confirm an hypothesis, and discriminate between competing hypotheses. These characterizations are in terms of prime implicates, and hence are implementable using assumption-based truth maintenance systems. Further, we provide results on the impact of tests within two specific hypothetical reasoning paradigms: consistency-based reasoning and abduction. Finally, we provide results on differential diagnosis for consistency-based and abductive reasoning. The results of this paper are relevant to a number of hypothetical reasoning tasks including diagnosis and active vision.

Preliminaries

We assume a fixed propositional language throughout. Σ will be a fixed sentence of the language, and will serve as the relevant background knowledge describing the system under analysis. For example, in the case of circuits, Σ might describe the individual circuit components, their normal input/output behaviour, their fault models, the topology of their interconnections, and the legal combinations of circuit inputs (e.g. [de Kleer and Williams, 1987], [Reiter, 1987]). We also assume a fixed set HYP of hypotheses. In the case where Σ describes a circuit, HYP might be the set of diagnoses which we currently hold for this device. How we arrived at the set HYP will be largely irrelevant for our purposes. HYP could be a set of abductive hypotheses [Poole, 1989], or the result of a consistency-based diagnostic procedure [de Kleer *et al.*, to appear], or any other conceivable form of hypothesis generation. Our one assumption about $H \in HYP$ is that H be a conjunction of literals of the propositional language.

Tests

Informally, the notion of a test provides for certain initial conditions which may be established by the tester, together with the specification of an observation whose outcome determines what the test conclusions are to be. For example, in circuit diagnosis the initial conditions of a test might be the provision of certain fixed circuit inputs, and the observation might be the resulting value

*on leave from the Alberta Research Council

†Fellow, Canadian Institute for Advanced Research

of a circuit output, or the value of an internal probe. In the medical setting, the initial conditions might involve performing a laboratory procedure like a blood test, and the observation might be the white cell count. In an active vision setting, the initial conditions might involve changing the camera angle or moving objects in the scene, and the observation might be some aspect of the corresponding image. We provide for a formal definition of a test by distinguishing a subset of literals of our propositional language, called the *achievable literals*. These will specify the initial conditions for a test. In addition, we require a distinguished subset of the propositional symbols of our language called the *observables*. These will specify the observations to be made as part of a test.

Definition 1 (Test) *A test is a pair (A, o) where A is a conjunction of achievable literals and o is an observable.*

A test specifies some initial condition A which the tester establishes, and an observation o whose truth value the tester is to determine.

Definition 2 (Outcome of a test) *The outcome of a test (A, o) is one of o, $\neg o$.*

In other words, as a result of performing the test (A, o) in the physical world, the truth value of o is observed. If o is observed to be true, the outcome of the test is o, otherwise $\neg o$.

Definition 3 (Confirmation, Refutation) *The outcome α of the test (A, o) confirms $H \in HYP$ iff $\Sigma \wedge A \wedge H$ is satisfiable, and $\Sigma \wedge A \models H \supset \alpha$. α refutes H iff $\Sigma \wedge A \wedge H$ is satisfiable, and $\Sigma \wedge A \models H \supset \neg\alpha$.*

At first, the requirement in this definition that $\Sigma \wedge A \wedge H$ be satisfiable might seem odd. However, not all conjunctions A of achievable literals will be legal initial conditions, for example simultaneously making a digital circuit input 0 and 1. Since Σ will encode constraints determining the legal initial conditions, we require that $\Sigma \wedge A$ be satisfiable. Moreover, hypothesis H could conceivably further constrain the possible initial conditions A permitted in a test. For example, the hypothesis that radioactivity has escaped within a reactor would prevent a test in which humans enter the reactor chamber. In such a case, Σ would include a formula of the form *radioactivity* $\supset \neg$*enter-chamber* so that $\Sigma \wedge$ *enter-chamber* $\wedge$ *radioactivity* would be unsatisfiable, in which case the very idea of a confirming or refuting outcome of such a test would be meaningless.

In general, a confirming outcome for H provides no deterministic information about H; we can neither accept nor reject H on the strength of the test outcome.[1] A refuting outcome for H, however, allows us to reject H as a possible hypothesis.

Definition 4 (Prime Implicate) *A prime implicate of a propositional formula Σ is a clause C such that $\Sigma \models C$, and for no proper subclause C' of C does $\Sigma \models C'$*

[1] Of course, H's probability may well increase as a result of a confirming outcome.

Theorem 1 *The outcome α of test (A, o) confirms (refutes) $H \in HYP$ iff*

1. *There is a prime implicate of Σ of the form $\neg A' \vee \neg H' \vee \alpha$ ($\neg A' \vee \neg H' \vee \neg\alpha$) where A' is a subconjunct of A and H' is a subconjunct of H, and*
2. *No prime implicate of Σ subsumes $\neg A \vee \neg H$.*

Proof: Suppose α confirms H. Then by definition, $\Sigma \wedge A \models H \supset \alpha$. Hence there is a prime implicate of Σ of the form $\neg A' \vee \neg H' \vee (\alpha)$ where A' and H' are subconjuncts of A and H respectively, and where the notation (α) indicates that the literal α may or may not be present in the clause. We prove that α is indeed present in the clause, in which case the desired result will be established. If in fact α is not present, then $\neg A' \vee \neg H'$ is a prime implicate of Σ, i.e. $\Sigma \wedge A' \wedge H'$ is unsatisfiable, in which case so is $\Sigma \wedge A \wedge H$, contradicting the assumption that α confirms H. To see that 2. must be true, assume on the contrary that some prime implicate of Σ subsumes $\neg A \vee \neg H$. This means that $\Sigma \wedge A \wedge H$ is unsatisfiable, which is impossible since α confirms H.
To prove the converse, suppose $\neg A' \vee \neg H' \vee \alpha$ is a prime implicate of Σ. Then $\Sigma \wedge A' \models H' \supset \alpha$, whence $\Sigma \wedge A \models H \supset \alpha$. Since condition 2. means that $\Sigma \wedge A \wedge H$ is satisfiable, it follows that α confirms H.
A similar argument establishes the theorem in the case of refutations.

Discriminating Tests

In this section we characterize those tests (A, o) which are guaranteed to discriminate an hypothesis space HYP, i.e. which will refute at least one hypothesis in HYP.

Definition 5 (Discriminating Tests) *A test (A, o) is a discriminating test for the hypothesis space HYP iff $\Sigma \wedge A \wedge H$ is satisfiable for all $H \in HYP$ and there exists H_i, $H_j \in HYP$ such that the outcome α of test (A, o) refutes either H_i or H_j, no matter what that outcome might be.*

In the case of uniformly distributed hypotheses, we would ideally like a discriminating test to refute half of the hypotheses in the hypothesis space, regardless of the test outcome. By definition, it must refute at least one hypothesis in the hypothesis space.

Definition 6 (Minimal Discriminating Tests) *A discriminating test (A, o) for the hypothesis space HYP is minimal iff for no proper subconjunct A' of A is (A', o) a discriminating test for HYP.*

Minimal discriminating tests preclude unnecessary initial conditions, for example unnecessary circuit inputs, laboratory tests, etc. Only those conditions necessary for producing the test outcome are invoked.

Theorem 2

1. *Suppose Σ has at least two prime implicates of the form $\neg A' \vee \neg H' \vee o$ and $\neg A'' \vee \neg H'' \vee \neg o$ where*
   (a) *H' and H'' are subconjuncts of H_i and H_j respectively, for some $H_i, H_j \in HYP$, and*
   (b) *No prime implicate of Σ subsumes $\neg A' \vee \neg A'' \vee \neg H$, for all $H \in HYP$.*

   Then $(A' \wedge A'', o)$ is a discriminating test for the hypothesis space HYP.

2. *Moreover, every minimal discriminating test can be obtained this way, i.e. if (A, o) is a minimal discriminating test for the hypothesis space HYP, then Σ has at least two prime implicates of the form $\neg A' \vee \neg H' \vee \pm o$ and $\neg A'' \vee \neg H'' \vee \mp o$ where*
   (a) *$A = A' \wedge A''$,*
   (b) *H' and H'' are subconjuncts of H_i and H_j respectively, $H_i, H_j \in HYP$, and*
   (c) *No prime implicate of Σ subsumes $\neg A \vee \neg H$ for all $H \in HYP$.*

Proof:

1. We prove the result in the case that o is the outcome of (A, o). A symmetrical proof applies when the outcome is $\neg o$. Since $\neg A' \vee \neg H' \vee o$ is a prime implicate of Σ, we have $\Sigma \wedge A' \models H' \supset o$. Thus $\Sigma \wedge A' \wedge A'' \models H_i \supset o$. Similarly, $\Sigma \wedge A' \wedge A'' \models H_j \supset \neg o$. Finally, since no prime implicate of Σ subsumes $\neg A' \vee \neg A'' \vee \neg H$ for all $H \in HYP$ then $\Sigma \wedge A' \wedge A'' \wedge H$ is satisfiable for all $H \in HYP$. Hence o confirms H_i and refutes H_j, so that $(A' \wedge A'', o)$ is a discriminating test for the hypothesis space HYP.

2. Suppose (A, o) is a minimal discriminating test for the hypothesis space HYP. Without loss of generality, assume that o is the outcome of (A, o), and that o confirms H_i and refutes H_j. Then by Theorem 1, Σ has two prime implicates of the form $\neg A' \vee \neg H' \vee o$ and $\neg A'' \vee \neg H'' \vee \neg o$, where A' and A'' are subconjuncts of A, and H' and H'' are subconjuncts of H_i and H_j respectively; moreover, no prime implicate of Σ subsumes $\neg A \vee \neg H$ for all $H \in HYP$. Hence, by part 1. of this theorem, $(A' \wedge A'', o)$ is a discriminating test for the hypothesis space HYP. Since $A' \wedge A''$ is a subconjunct of A, and since (A, o) is a minimal discriminating test for the hypothesis space HYP, $A = A' \wedge A''$.

An interesting special case of Theorem 2 arises when there are no initial conditions, i.e. when a simple system observation is to be made, without establishing initial conditions for the test. This is the case $A = true$.

Corollary 1 *Suppose that $\Sigma \wedge H$ is satisfiable for all $H \in HYP$.* [2] *Then $(true, o)$ is a discriminating test (and hence a minimal discriminating test) for the hypothesis space HYP iff Σ has at least two prime implicates of the form $\neg H' \vee \pm o$ and $\neg H'' \vee \mp o$ where H' and H'' are subconjuncts of H_i and H_j respectively, $H_i, H_j \in HYP$.*

[2] Notice that this will be the normal case. No one would entertain an hypothesis which is inconsistent with the background theory Σ.

In [Sattar and Goebel, 1989], Sattar and Goebel provide a mechanism within the Theorist system [Poole *et al.*, 1987] for recognizing so-called *crucial literals* which provides a basis for performing discriminating tests without initial conditions. The above corollary is an abstract characterization of their method, with o playing the role of their crucial literal. More preliminary work on crucial literals for discriminating competing hypotheses can be found in [Seki and Takeuchi, 1985] and [Shapiro, 1981]. Genesereth's work on the DART system [Genesereth, 1984] is also similar is spirit. DART was capable of designing tests by a process (called *residue resolution*) very like the generation of prime implicates. The above results can be viewed as a systematic exploration of some of the ideas embodied in the DART program.

Relevant Tests

In many instances we will not have discriminating tests. Our concern here is characterizing *relevant tests*; those tests (A, o) which have the *potential* to discriminate an hypothesis space HYP but which cannot be guaranteed to do so. Given a particular outcome α, relevant tests will refute a subset of the hypotheses in the hypothesis space HYP, but may not refute any hypotheses if we do not observe α. Since we have no guarantee a priori of the outcome of a test, these tests are not guaranteed to discriminate an hypothesis space. In the section on tests for hypothetical reasoning we will show that relevant tests are guaranteed to discriminate a space of abductive hypotheses.

Definition 7 (Relevant Tests) *A test (A, o) is a relevant test for the hypothesis space HYP iff $\Sigma \wedge A \wedge H$ is satisfiable for all $H \in HYP$ and the outcome α of test (A, o) either confirms a subset of the hypotheses in HYP or refutes a subset.*

Note, by definition, that every discriminating test is a relevant test. Informally, a relevant test is one that produces a particular outcome when certain hypotheses drawn from HYP are true, but does not produce that outcome when other hypotheses drawn from HYP are true. Consider a simple medical diagnosis problem where we suspect that a patient is suffering from either mumps, measles, chicken pox or the flu. Both the hypothesis that the patient has measles and the hypothesis that the patient has chicken pox, infer the observation of red spots. However, neither the hypothesis that the patient has mumps or the hypothesis that the patient has the flu infer anything about the existence or lack of existence of red spots. As a result, the outcome of a test to observe red spots will only provide discriminatory information if we observe red spots to be false. In such a case we can refute both the chicken pox and the measles. However, if we observe red spots to be true, we are unable to reject any of the four hypotheses.

Definition 8 (Minimal Relevant Tests)
A relevant test (A, o) for the hypothesis space HYP is minimal iff for no proper subconjunct A' of A is (A', o) a relevant test for HYP.

Again, minimal relevant tests preclude unnecessary initial conditions.

Theorem 3

1. *Suppose Σ has at least one prime implicate of the form $\neg A \vee \neg H' \vee o$, where H' is a subconjunct of some $H_i \in HYP$. Further suppose there exists $H_j \in HYP$ such that Σ has no prime implicate of the form $\neg A' \vee \neg H'' \vee o$, where A' and H'' are subconjuncts of A and H_j, respectively. Finally, suppose no prime implicate of Σ subsumes $\neg A \vee \neg H$ for all $H \in HYP$. Then (A, o) is a relevant test for the hypothesis space HYP.*
2. *Moreover, every minimal relevant test can be obtained this way, i.e. if (A, o) is a minimal relevant test for the hypothesis space HYP, then Σ has a prime implicate of the form $\neg A \vee \neg H' \vee o$ where,*
   - *(a) there exists $H_j \in HYP$ such that there is no prime implicate of the form $\neg A' \vee \neg H'' \vee o$, H'' and A' are subconjuncts of H_j and A respectively; and*
   - *(b) no prime implicate of Σ subsumes $\neg A \vee \neg H$ for all $H \in HYP$.*

Proof:

1. We prove the result in the case that o is the outcome of (A, o). A symmetrical proof applies when the outcome is $\neg o$. Since $\neg A \vee \neg H' \vee o$ is a prime implicate of Σ, we have $\Sigma \wedge A \models H' \supset o$. Furthermore, since no prime implicate of Σ subsumes $\neg A \vee \neg H_i$, then $\Sigma \wedge A \wedge H_i$ is satisfiable. Hence, if the outcome α of test (A, o) is o, then α confirms $H_i \in HYP$. Since $\neg A' \vee \neg H'' \vee o$ is not a prime implicate of Σ, there is some $H_j \in HYP$ for which $\Sigma \wedge A \not\models H'' \supset o$. α does not confirm H_j. Finally since no prime implicate of Σ subsumes $\neg A \vee \neg H$ for all $H \in HYP$, then $\Sigma \wedge A \wedge H$ is satisfiable for all $H \in HYP$. Thus the test (A, o) is a relevant test for the hypothesis space HYP.
2. Suppose (A, o) is a minimal relevant test for the hypothesis space HYP. Without loss of generality, assume that o is the outcome of (A, o), and that o confirms H_i and does not confirm H_j. Then by Theorem 1, Σ has a prime implicate of the form $\neg A \vee \neg H' \vee o$, where H' is a subconjunct of $H_i \in HYP$, and no prime implicate of Σ subsumes $\neg A \vee \neg H_i$. Moreover, for some $H_j \in HYP$, Σ does not have a prime implicate of the form $\neg A' \vee \neg H'' \vee o$, where A' and H'' are subconjuncts of A and H_j respectively. Finally since no prime implicate of Σ subsumes $\neg A \vee \neg H$ for all $H \in HYP$, then $\Sigma \wedge A \wedge H$ is satisfiable for all $H \in HYP$. Hence, by part 1. of this theorem, (A, o) is a relevant test for the hypothesis space HYP.

Why Prime Implicates?

The characterizing theorems of the previous section are in terms of the prime implicates $PI(\Sigma)$ of Σ. Thus Theorem 1 informs us how to "read off", from $PI(\Sigma)$, all hypotheses confirmed or refuted by the outcome of a given test. Alternatively, Theorem 1 informs us how to determine all tests whose outcomes can confirm or refute a given hypothesis. Similarly, Theorem 2 can be used to determine all pairs (H_i, H_j) (or more) of hypotheses for which a given test (A, o) is guaranteed to be a discriminating test. Theorem 2 can also be used to determine all minimal discriminating tests for a given pair (H_i, H_j) (or more) of hypotheses. Finally, Theorem 3 tells us how to detect minimal relevant tests for a space of hypotheses by examining $PI(\Sigma)$. It additionally indicates which hypotheses could potentially be confirmed or refuted by a given test (A, o). Provided $PI(\Sigma)$ has already been computed, all these tasks are straightforward and computationally attractive. Alas, as is well known, computing $PI(\Sigma)$ is computationally intractable ([Bylander *et al.*, 1989], [Selman and Levesque, 1990]), and not only because there may be exponentially many prime implicates. As it happens, the principal task of an assumption-based truth maintenance system is the computation of certain prime implicates of a background theory Σ [Reiter and de Kleer, 1987]. Despite the high complexity associated with the computation of prime implicates, ATMSs are very frequently used as implementation tools in abductive and diagnostic reasoning systems. Therefore, in those cases where an ATMS is providing the underlying reasoning service, the results on the design of tests of the previous section are especially relevant. In effect, the ATMS will have already performed all of the preliminary work – namely the calculation of the prime implicates – necessary for applying the results of the previous section. We obtain the benefits of this analysis of tests as a side effect of the ATMS calculations when achievable literals are encoded as assumptions.

ATMS *assumptions* encode the distinguished literals from which hypotheses are generated. Achievable literals may be encoded as additional assumptions. An observable o is a *datum* of an ATMS *node*. The *label* of the node representing o contains the set of *environments* in which o is true. Thus (A, o) is a test for H if one of the environments in the label of o contains the set of literals from which A', H' (subconjuncts of A and H) are generated. A test (A, o) discriminates two hypotheses H_1 and H_2 if nodes for o and $\neg o$ exist such that (A, o) is a test for H_1 and $(A, \neg o)$ is an test for H_2. Tests may be selected by inspecting the labels of the nodes of observable data.

Tests for Hypothetical Reasoning

In the previous sections we characterized the notion of a test and demonstrated how it might be realized in an implementation such as the ATMS. In what follows we examine the role of tests within the context of two specific hypothetical reasoning paradigms: consistency-based reasoning and abduction [de Kleer *et al.*, to appear]. In particular, we show that the discriminatory power of a test outcome is contingent upon the hypothetical reasoning paradigm.

To this end, we must assume a distinguished finite subset $\mathcal{H} = \{h_1, \ldots, h_n\}$ of propositional symbols which will function as the primitive hypotheses. Let $conj(\mathcal{H})$ be the set of all conjunctions of the form $l_1 \wedge \cdots \wedge l_n$ where l_i is a literal and h_i is the propositional symbol mentioned by l_i. For example, if $\mathcal{H} = \{h_1, h_2\}$, then $conj(\mathcal{H}) = \{h_1 \wedge h_2, h_1 \wedge \neg h_2, \neg h_1 \wedge h_2, \neg h_1 \wedge \neg h_2\}$. Each hypothesis commits to the truth or falsity of every primitive hypothesis in $\mathcal{H}$. $conj(\mathcal{H})$ is analogous to the expressions $[\bigwedge_{c \in C_p} AB(c)] \wedge [\bigwedge_{c \in C_n} \neg AB(c)]$ in [de Kleer *et al.*, to appear] from which diagnoses are to be drawn.

Although the propositions in this section are defined in terms of the hypothesis space $conj(\mathcal{H})$, the results are also true for minimal and kernel, abductive and consistency-based hypothesis spaces.

Definition 9 (Consistency-based Hypotheses) *A consistency-based hypothesis for Σ and outcome α of the test (A, o) is any $H \in conj(\mathcal{H})$ such that $\Sigma \wedge A \wedge H \wedge \alpha$ is satisfiable.*

This definition follows from [de Kleer *et al.*, to appear].

Proposition 1 (Consistency-based Tests) *The outcome α of a relevant test (A, o) discriminates a space of consistency-based hypotheses HYP iff the outcome refutes some hypotheses in HYP.*

As a consequence, a relevant test is not guaranteed to discriminate a consistency-based hypothesis space.

Definition 10 (Abductive Hypotheses) *An abductive hypothesis for Σ and outcome α of the test (A, o) is any $H \in conj(\mathcal{H})$ such that $\Sigma \wedge A \wedge H \models \alpha$ and $\Sigma \wedge A \wedge H$ is satisfiable.*

This definition follows from [de Kleer *et al.*, to appear]. The outcome of a relevant test has a substantially different impact on a space of abductive hypotheses as illustrated by the following example.

Example Let Σ be the sentence $(h_1 \supset a) \wedge (h_2 \supset b)$, and suppose that the hypotheses are drawn from the vocabulary $\{h_1, h_2\}$. Finally, suppose the initial space of hypotheses – say as a result of the observation *true* – is

$\{h_1 \wedge h_2, h_1 \wedge \neg h_2, \neg h_1 \wedge h_2, \neg h_1 \wedge \neg h_2\}$.

After explaining the outcome a of the test $(true, a)$, the set of abductive hypotheses HYP is

$$\{h_1 \wedge h_2, h_1 \wedge \neg h_2.\}$$

But the outcome a refutes none of the original abductive hypotheses.

Abduction demands that $\Sigma \wedge H \models \alpha$. Hence, by definition, hypotheses that confirm α are abductive hypotheses. However, all other hypotheses that are consistent with Σ but for which $\Sigma \wedge H \not\models \alpha$, are not abductive hypotheses. Thus, a test outcome that does not confirm an hypothesis, whether it explicitly refutes it or not, causes that hypothesis to be rejected.

Proposition 2 (Abductive Tests) *The outcome α of a relevant test (A, o) is guaranteed to discriminate the space of abductive hypotheses HYP .*

Propositions 1 and 2 demonstrate that by exploiting the fact that abductive hypotheses must *explain* test outcomes rather than just be *consistent* with those outcomes, we may discriminate abductive hypothesis spaces with any relevant test, regardless of its outcome. Thus, in an abductive setting, relevant tests and discriminating tests are equally desirable. The strategy for selection of tests to discriminate a space of hypotheses is impacted accordingly.

Example (continued) Following the outcome a of test $(true, a)$, the abductive hypothesis space HYP was: $\{h_1 \wedge h_2, h_1 \wedge \neg h_2\}$.

Assume we now perform the test $(true, b)$ and observe the outcome $\neg b$. Since the outcome $\neg b$ refutes h_2, we would expect to reject the hypothesis $h_1 \wedge h_2$. However, with the above definition of abductive hypotheses, we will also reject the hypothesis $h_1 \wedge \neg h_2$ because $\Sigma \wedge h_1 \wedge \neg h_2 \not\models \neg b$.

This illustrates an interesting point regarding the behaviour of abduction which is relevant to testing in particular and to abduction in general. Abductive reasoning requires that every observation which we might wish to explain be encoded explicitly in Σ. Unexpected or irrelevant observations (such as for example, that a patient being diagnosed has curly red hair) will cause all abductive hypotheses to be rejected because none of them can explain the observation. This seemingly undesirable behaviour is avoided in a testing environment because the only observations provided to the system are those that Σ generates as *test* observations. In other abductive reasoning environments, such rejected hypotheses might be tagged as partial explanations, but this can complicate the comparison of final hypotheses. This behaviour has ramifications for negative observations which, unless explicitly encoded, will also cause the rejection of hypotheses. Characterizing abduction as generalized stable models and implementing negation as failure [Preist and Eshghi, 1992] is a solution to this problem.

Differential Diagnosis

To this point, discussion has been limited to the execution of a single test and its impact on a space of hypotheses. Our objective here is to characterize differential diagnosis (DD) for consistency-based and abductive hypothesis spaces. Differential diagnosis is one of several sequential diagnosis or sequential hypothetical reasoning strategies.

The intuitive notion of differential diagnosis as described by Ledley and Lusted [Ledley and Lusted, 1959] is this: Given a set of potential diagnoses, a sequence of tests may be performed to iteratively reject diagnoses *without the need for subsequent diagnosis generation steps*. Following each test, the resulting set of hypotheses contains all and only the hypotheses to be entertained in further hypothetical reasoning.

The Differential Diagnosis Principle (DDP) Given HYP, Σ, (A, o) and α as above, the differential diagnosis principle is that

$$\{H \in HYP \mid \alpha \textit{ does not eliminate } H\}$$

the set of hypotheses for $\Sigma \wedge A \wedge \alpha$ is a subset of HYP.

Notice that the new background theory is $\Sigma \wedge A \wedge \alpha$, reflecting the new background knowledge resulting from the performance of the test.

The correctness of DDP, and further, the criteria by which α rejects hypotheses depend crucially on the nature of the initial hypothesis set HYP. For example, DDP does not apply when HYP is taken to be the set of minimal or kernel diagnoses, whether consistency-based or abductive, as defined in [de Kleer *et al.*, to appear]. In both these cases, test outcomes do not simply result in the pruning of the hypothesis space, but may require the generation of new hypotheses.

In what follows, we characterize differential diagnosis for consistency-based hypotheses and for abductive hypotheses [de Kleer *et al.*, to appear]. Furthermore, we show that the results for the space of abductive hypotheses also hold for consistency-based hypotheses when we restrict Σ to a closed simple causal theory.

Theorem 4 (Consistency-based DD) *Suppose HYP is the set of all consistency-based hypotheses for Σ, and let α be the outcome of the test $(true, o)$. Then*

$$NEWHYP = \{H \in HYP \mid \alpha \textit{ does not refute } H\}$$

is the set of consistency-based hypotheses for $\Sigma \wedge \alpha$. [3]

Proof: Let $H \in conj(\mathcal{H})$. We must prove that $H \in NEWHYP$ iff $\Sigma \wedge \alpha \wedge H$ is satisfiable. Suppose $H \in NEWHYP$. Then α does not refute H, which is to say, $\Sigma \not\models H \supset \neg\alpha$, i.e. $\Sigma \wedge \alpha \wedge H$ is satisfiable, so that H is a consistency-based hypothesis for $\Sigma \wedge \alpha$. Conversely, suppose $\Sigma \wedge \alpha \wedge H$ is satisfiable. Then $\Sigma \not\models H \supset \neg\alpha$, i.e. α does not refute H. Moreover, $\Sigma \wedge H$ is satisfiable, so that $H \in HYP$. Hence $H \in NEWHYP$.

Theorem 5 (Abductive DD) *Suppose HYP is the set of all abductive hypotheses for Σ, and let α be the outcome of the test $(true, o)$. Then*

$$NEWHYP = \{H \in HYP \mid \alpha \textit{ confirms } H\}$$

is the set of abductive hypotheses for $\Sigma \wedge \alpha$.

Proof: Let $H \in conj(\mathcal{H})$. We must prove that $H \in NEWHYP$ iff $\Sigma \wedge H$ is satisfiable and $\Sigma \wedge H \models \alpha$. Suppose $H \in NEWHYP$. Then α confirms H, which is to say, $\Sigma \models H \supset \alpha$, i.e. $\Sigma \wedge H$ is satisfiable and $\Sigma \wedge H \models \alpha$, so that H is an abductive hypothesis for $\Sigma \wedge \alpha$. Conversely, suppose $\Sigma \wedge H$ is satisfiable and $\Sigma \wedge H \models \alpha$. Then $\Sigma \models H \supset \alpha$, i.e. α confirms H. Moreover, $\Sigma \wedge H$ is satisfiable, so that $H \in HYP$. Hence $H \in NEWHYP$.

In the following section we see that by restricting the form of Σ, we can acquire the same result as Theorem 6.2 for consistency-based hypotheses.

Consistency-based DD of Causal Theories

Poole [Poole, 1988] and Konolige [Konolige, to appear] have studied consistency-based and abductive diagnosis for what Konolige refers to as *simple causal theories*. They have shown that the minimal abductive diagnoses for a simple causal theory are identical to the minimal consistency-based diagnoses for the Clark completion [Clark, 1978] of a simple causal theory. In keeping with the spirit of that work, we characterize differential diagnosis for *closed simple causal theories*, which we show to be equivalent to abductive differential diagnosis.

Definition 11 (Simple Causal Theory) *Let $\mathcal{L}$ be a propositional language. A simple causal theory is a tuple (C, E, Σ) where*

1. *C, a set of atomic sentences of $\mathcal{L}$, is the set of causes.*
2. *E, a set of atomic sentences of $\mathcal{L}$, is the set of effects we might observe and whose causes we seek.*
3. *Σ, a set of sentences of $\mathcal{L}$, is the domain theory, containing information about the relation between causes and effects. The sentences of Σ have the form $\mathcal{C} \supset e$ where $e \in E$ and $\mathcal{C}$ is a conjunction of literals whose propositional symbols are causes.*

This definition follows from [Konolige, to appear].

Definition 12 (Closed Simple Causal Theory) *Let (C, E, Σ) be a simple causal theory over a propositional language with Σ a set of nonatomic definite clauses whose directed graph is acyclic. Then we define Σ^*, the closed simple causal theory, to be Σ augmented by the Clark completion [Clark, 1978] of Σ.*

The above definition follows from Theorem 1 in [Konolige, to appear].

Theorem 6 (Consistency-based DD of Σ^*) *Suppose that (C, E, Σ) is a simple causal theory, that HYP is the set of all consistency-based hypotheses for Σ^*, and that α is the outcome of the test $(true, o)$, where $o \in E$. Then*

$$NEWHYP = \{H \in HYP \mid \alpha \textit{ confirms } H\}$$

is the set of consistency-based hypotheses for $\Sigma^* \wedge \alpha$.

Proof: Let $H \in conj(\mathcal{H})$. We must prove that $H \in NEWHYP$ iff $\Sigma^* \wedge \alpha \wedge H$ is satisfiable. Suppose $H \in NEWHYP$. Then α confirms H, so $\Sigma^* \wedge H$

[3] Notice that the theorem is stated only for simple tests of the form $(true, o)$, not for (A, o) for arbitrary initial conditions A. The general case is somewhat problematic; we shall discuss it in the next section.

is satisfiable and $\Sigma^* \wedge H \models \alpha$. Hence $\Sigma^* \wedge H \wedge \alpha$ is satisfiable, so that H is a consistency-based hypothesis for $\Sigma^* \wedge \alpha$. Conversely, suppose $\Sigma^* \wedge H \wedge \alpha$ is satisfiable. Then $\Sigma^* \not\models H \supset \neg\alpha$. We prove that $\Sigma^* \models H \supset \alpha$ or $\Sigma^* \models H \supset \neg\alpha$, from which the result will follow. To that end, notice that in view of the fact that Σ^* is the Clark completion of Σ, $\Sigma^* \models \alpha \equiv B$ where B is a sentence, all of whose propositional atoms are in $\mathcal{H}$. Since $H \in conj(\mathcal{H})$, every atom mentioned by B is mentioned by H, so that $\models H \supset B$ or $\models H \supset \neg B$. Hence $\Sigma^* \models H \supset \alpha$ or $\Sigma^* \models H \supset \neg\alpha$.

The restriction to a closed simple causal theory is limiting. Konolige [Konolige, to appear] discusses the conditions under which closure axioms may be consistently added to a theory. A significant benefit of closure axioms is that they enable explanations of test outcomes to be generated deductively.

The results of Theorems 4, 5 and 6 may be applied to strategies for the selection of tests. For example, in order to isolate a unique hypothesis from a space of consistency-based hypotheses, differential diagnosis must select a sequence of tests to refute all of the hypotheses but one. By contrast, a unique abductive hypothesis may be isolated either by selecting a sequence of tests to refute all other hypotheses, or simply by selecting one or more tests which uniquely confirm a hypothesis. In the case where the hypotheses are not equally likely, this is a particularly attractive test selection strategy.

Discussion

As noted above, the differential diagnosis theorems were proven for tests of the form $(true, o)$. Differential diagnosis for arbitrary tests (A, o) is more difficult to characterize because the realization of initial conditions A could have side effects in the world which would change the truth value of previous observations. For example, if we execute a test to biopsy a tumor and A involves removal of the tumor, then a side effect will be that the tumor is no longer present, which would contradict any previous observation relating the existence and location of the tumor.

In order to characterize differential diagnosis for arbitrary tests (A, o), time must be taken into account to index test conditions and observations, and to reason about them accordingly. Although the characterization of testing provided in this paper is sufficient for many hypothetical reasoning tasks, other tasks require a more sophisticated formalism where we characterize test conditions as *actions* in situation calculus with preconditions and postconditions. (Provan and Poole [Provan and Poole, 1991]) and (Webber et al. [Webber *et al.*, 1990]) have emphasized the importance of including treatment in the diagnostic process; treatments may be similarly encoded as actions. In fact, some treatments (such as replacing the battery in an electronic device) play a dual role as treatment and test.

Test selection traditionally involves utility measures such as time, cost, and information gain (e.g. deKleer and Williams [de Kleer and Williams, 1987]). Planning also plays a role in test design [Webber *et al.*, 1990]. There are at least two distinct objectives of planning in the diagnostic setting. One is to achieve some *state of the world*, as for example planning a suitable sequence of steps in order to insert a measuring probe in some device, or more globally planning a sequence of tests and treatments to ascertain and eradicate undesirable system behaviour. The other objective is to achieve a suitable *state of knowledge*, for example taking a person's temperature in order to *know whether* she has a fever, or more globally, planning a sequence of tests and treatments to reach a state of knowledge where only one hypothesis remains or where a particular hypothesis has been eliminated. These two planning objectives are quite different. Both may be modeled in the situation calculus [McCarthy and Hayes, 1969], but the latter requires formalization in an epistemic logic, along the lines of [Moore, 1985]. We are currently investigating the extension of the propositional logic formalism of this paper to a situation calculus planning formalism for these and other related problems in hypothetical reasoning.

Acknowledgements

The authors would like to thank David Poole, Abdul Sattar and Ron Rymon for helpful comments on an earlier draft. Additionally, the first author would like to thank Michael Gruninger, Javier Pinto and Dale Schuurmans for their comments and friendship throughout the writing of this paper.

References

[Bylander *et al.*, 1989] T. Bylander, D. Allemang, M.C. Tanner, and J.R. Josephson. Some results concerning the computational complexity of abduction. In R. Brachman, H.J. Levesque, and R. Reiter, editors, *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning (KR'89)*, pages 44–54. Morgan Kaufmann Publishers, Inc., 1989.

[Clark, 1978] K.L. Clark. Negation as failure. In H. Gallaire and J. Minker, editors, *Logic and Data Bases*, pages 292–322. Plenum Press, New York, 1978.

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

[de Kleer *et al.*, to appear] J. de Kleer, A.K. Mackworth, and R. Reiter. Characterizing diagnoses and systems. *Artificial Intelligence*, to appear.

[Genesereth, 1984] M.R. Genesereth. The use of design descriptions in automated diagnosis. *Artificial Intelligence*, 24:411–436, 1984.

[Konolige, to appear] K. Konolige. Abduction vs. closure in causal theories. *Artificial Intelligence*, to appear.

[Ledley and Lusted, 1959] R.S. Ledley and L.B. Lusted. Reasoning foundations of medical diagnosis. *Science*, 130(3366):9–21, 1959.

[McCarthy and Hayes, 1969] J. McCarthy and P. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, Scotland, 1969.

[Moore, 1985] R.C. Moore. A formal theory of knowledge and action. In Jerry B. Hobbs and Robert C. Moore, editors, *Formal Theories of the Commonsense World*, chapter 9, pages 319–358. Ablex Publishing Corp., Norwood, New Jersey, 1985.

[Poole *et al.*, 1987] D. Poole, R.G. Goebel, and R. Aleliunas. Theorist: a logical reasoning system for defaults and diagnosis. In N. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 331–352. Springer Verlag, 1987.

[Poole, 1988] D. Poole. Representing knowledge for logic-based diagnosis. In *Proceedings of the Fifth Generation Computer Systems Conference (FGCS'88)*, pages 1282–1290, 1988.

[Poole, 1989] D. Poole. Explanation and prediction: an architecture for default and abductive reasoning. *Computational Intelligence*, 5:97–110, 1989.

[Preist and Eshghi, 1992] C. Preist and K Eshghi. Consistency-based and abductive diagnoses as generalised stable models. In *Proceedings of the Fifth Generation Computer Systems Conference (FGCS'92)*, page to appear, 1992.

[Provan and Poole, 1991] G. Provan and D. Poole. The utility of consistency-based diagnostic techniques. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 461–472. Morgan Kaufmann Publishers, Inc., 1991.

[Reiter and de Kleer, 1987] R. Reiter and J. de Kleer. Foundations for assumption-based truth maintenance systems: Preliminary report. In *Proceedings of the National Conference on Artificial Intelligence*, pages 183–188, 1987.

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

[Sattar and Goebel, 1989] A. Sattar and R. Goebel. Using crucial literals to select better theories. Technical Report TR 89-27, Department of Computing Science, University of Alberta, 1989.

[Seki and Takeuchi, 1985] A. Seki and A. Takeuchi. An algorithm for finding a query which discriminates competing hypotheses. Technical Report TR 143, Institute for New Generation Computer Technology, Tokyo, Japan, 1985.

[Selman and Levesque, 1990] B. Selman and H.J. Levesque. Abductive and default reasoning: a computational core. In *Proceedings of the National Conference on Artificial Intelligence*, pages 343–348, 1990.

[Shapiro, 1981] E. Shapiro. Inductive inference of theories from fact. Technical Report TR 192, Department of Computer Science, Yale University, 1981.

[Webber *et al.*, 1990] B. Webber, R. Clarke, M. Niv, R. Rymon, and M. Milagros Ibanez. TraumAID: reasoning and planning in the initial definitive management of multiple injuries. Technical Report MS-CIS-90-50, Department of Computer and Information Science, University of Pennsylvania, 1990.