On Experiments for Hypothetical Reasoning

Sheila McIlraith
and
Raymond Reiter*
Department of Computer Science
University of Toronto
Toronto, Canada M5S 1A4
email: mcilrait@ai.toronto.edu, reiter@ai.toronto.edu
fax: 416-978-1931

January 13, 1992

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 experiments – how they can be designed, and what conclusions can be drawn about the hypotheses in HYP as a result of performing experiments. Specifically, we define the concept of an experiment and the concept of the outcome of an experiment. We characterize those experiments 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. Finally, we provide some results on differential diagnosis in the cases of consistency and abductive-based diagnosis.

1 Introduction

In the AI literature on hypothetical reasoning there are relatively few results on the design of experiments for discriminating between competing hypotheses, or on the conclusions one may draw from the outcome of an experiment. There are exceptions, of course. Among these are de Kleer and Williams [4] who provide a probabilistic analysis to decide what measurement to take next. The DART system of Genesereth [5] was capable of proposing circuit inputs and observations to be made in order to confirm or refute a possible diagnosis. TraumAID (Webber et al. [18]) 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 experiments in hypothetical reasoning. This paper is a first step in this direction.

Our concern in this paper is how experiments provide information about the current space of hypotheses. Specifically, we define the concept of an experiment and the concept of the outcome of an experiment. We characterize those experiments whose outcomes refute or confirm an hypothesis, and discriminate between competing hypotheses. These characterizations are in terms

*Fellow, Canadian Institute for Advanced Research

of prime implicates, and hence are implementable using assumption-based truth maintenance systems. Finally, we provide some results on differential diagnosis in the case of consistency and abductive-based diagnosis.

2 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. deKleer and Williams [4], Reiter [14]). 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 [12]), or the result of a consistency-based diagnostic procedure (deKleer, Mackworth and Reiter [3]), 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.

3 Experiments

Informally, the notion of an experiment provides for certain initial conditions which may be established by the experimenter, together with the specification of an observation whose outcome determines what the experimental conclusions are to be. For example, in circuit diagnosis the initial conditions of an experiment might be the provision of certain fixed circuit inputs, and the observation might be the resulting value 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. We provide for a formal definition of experiment by distinguishing a subset of literals of our propositional language, called the *achievable literals*. These will specify the initial conditions for an experiment. 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 an experiment.

Definition 3.1 (Experiment) *An experiment is a pair (A, o) where A is a conjunction of achievable literals and o is an observable.*

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

Definition 3.2 (Outcome of an Experiment) *The outcome of an experiment (A, o) is one of o, $\neg o$.*

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

Definition 3.3 (Confirmation, Refutation) *The outcome α of the experiment (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 an experiment. For example, the hypothesis that radioactivity has escaped within a reactor would prevent an experiment in which humans enter the reactor chamber. In such a case, Σ would include a formula of the form $radioactivity \supset \neg enter\text{-}chamber$ so that $\Sigma \wedge radioactivity \wedge enter\text{-}chamber$ would be unsatisfiable, in which case the very idea of a confirming or refuting outcome of such an experiment 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 experimental outcome.[1] A refuting outcome for H, however, allows us to reject H as a possible hypothesis.

Definition 3.4 *A prime implicate of a propositional formula Σ is a clause C such that*

1. $\Sigma \models C$, *and*
2. *For no proper subclause C' of C does $\Sigma \models C'$*

Theorem 3.1 *The outcome α of experiment (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, ie. $\Sigma \wedge A' \wedge H'$ 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.

3.1 Discriminating Experiments

Our concern here is characterizing those experiments (A, o) which, no matter what their outcome, are guaranteed to refute one of two competing hypotheses $H_1, H_2 \in HYP$.

Definition 3.5 (Discriminating Experiments) *An experiment (A, o) is a discriminating experiment for (H_1, H_2) iff its outcome refutes exactly one of H_1, H_2, no matter what that outcome might be.*

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

Definition 3.6 (Minimal Discriminating Experiments) *A discriminating experiment (A, o) for (H_1, H_2) is minimal iff for no proper subconjunct A' of A is (A', o) a discriminating experiment for (H_1, H_2).*

Minimal discriminating experiments preclude unnecessary initial conditions, for example unnecessary circuit inputs, laboratory tests, etc. Only those initial conditions necessary for discriminating H_1 and H_2 are invoked.

Theorem 3.2

1. *Suppose 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) H' and H'' are subconjuncts of H_1 and H_2 respectively, and

   (b) No prime implicate of Σ subsumes $\neg A' \vee \neg A'' \vee \neg H_1$ or $\neg A' \vee \neg A'' \vee \neg H_2$.

   Then $(A' \wedge A'', o)$ is a discriminating experiment for (H_1, H_2).

2. *Moreover, every minimal discriminating experiment can be obtained this way, i.e. if (A, o) is a minimal discriminating experiment for (H_1, H_2), then Σ has 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_1 and H_2 respectively, and

   (c) No prime implicate of Σ subsumes $\neg A \vee \neg H_1$ or $\neg A \vee \neg H_2$.

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 $\Sigma \wedge A' \wedge A'' \models H_1 \supset o$. Similarly, $\Sigma \wedge A' \wedge A'' \models H_2 \supset \neg o$. Finally, since no prime implicate of Σ subsumes $\neg A' \vee \neg A'' \vee \neg H_1$ or $\neg A' \vee \neg A'' \vee \neg H_2$, both $\Sigma \wedge A' \wedge A'' \wedge H_1$ and $\Sigma \wedge A' \wedge A'' \wedge H_2$ are satisfiable. Hence o confirms H_1 and refutes H_2, so that $(A' \wedge A'', o)$ is a discriminating experiment for (H_1, H_2).

2. Suppose (A, o) is a minimal discriminating experiment for (H_1, H_2). Without loss of generality, assume that o is the outcome of (A, o), and that o confirms H_1 and refutes H_2. Then by Theorem 3.1, 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_1 and H_2 respectively; moreover, no prime implicate of Σ subsumes $\neg A \vee \neg H_1$ or $\neg A \vee \neg H_2$. Hence, by part 1. of this theorem, $(A' \wedge A'', o)$ is a discriminating experiment for (H_1, H_2). Since $A' \wedge A''$ is a subconjunct of A, and since (A, o) is a minimal discriminating experiment for (H_1, H_2), $A = A' \wedge A''$.

An interesting special case of Theorem 3.2 arises when there are no initial conditions, i.e. when a simple system observation is to be made, without establishing initial conditions for the experiment. This is the case $A = true$.

Corollary 3.1 *Suppose $H_1, H_2 \in HYP$, and that $\Sigma \wedge H_1$ and $\Sigma \wedge H_2$ are satisfiable.*[2] *Then $(true, o)$ is a discriminating experiment (and hence a minimal discriminating experiment) for (H_1, H_2) iff Σ has 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_1 and H_2 respectively.*

In [16], Sattar and Goebel provide a mechanism within the Theorist system (Poole [13]) for recognizing so-called *crucial literals* which provide a basis for performing discriminating experiments without initial conditions. The above corollary is an abstract characterization of their method, with o playing the role of their crucial literal.

The only other work of which we are aware which is similar in spirit to our results on experiments is that of Genesereth for the DART system [5]. DART was capable of designing experiments 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.

4 Why Prime Implicates?

The characterizing theorems of the previous section are in terms of the prime implicates $PI(\Sigma)$ of Σ. Thus Theorem 3.1 informs us how to "read off", from $PI(\Sigma)$, all hypotheses confirmed or refuted by the outcome of a given experiment. Alternatively, Theorem 3.1 informs us how to determine all experiments whose outcomes can confirm or refute a given hypothesis. Similarly, Theorem 3.2 can be used to determine all pairs (H_1, H_2) of hypotheses for which a given experiment (A, o) is guaranteed to be a discriminating experiment. Theorem 3.2 can also be used to determine all minimal discriminating experiments for a given pair (H_1, H_2) of hypotheses. 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, Allemang, Tanner and Josephson [1], Selman and Levesque [17]), 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 all the prime implicates of a background theory Σ (Reiter and de Kleer [15]). 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 experiments 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 experiments as a free side effect of the ATMS calculations.

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 an experiment 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. An experiment (A, o) discriminates two hypotheses H_1 and H_2 if nodes for o and $\neg o$ exist such that (A, o) is an experiment for H_1 and $(A, \neg o)$ is an experiment for H_2. Experiments may be selected by inspecting the labels of the nodes of observable data.

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

5 Differential Diagnosis

The intuitive notion of differential diagnosis as described by Ledley and Lusted [7] is this: Given a set of potential diagnoses, a sequence of experiments may be performed to iteratively reject diagnoses *without the need for subsequent diagnosis generation steps.* Following each experiment, 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 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 experiment.

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 as defined in (deKleer, Mackworth and Reiter [3]). In both these cases, experiment 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 (deKleer, Mackworth and Reiter [3]) and for abductive hypotheses (Poole [12]). In keeping with intuition and with Popper's notion of falsifiability [10], we show that consistency-based hypotheses may be rejected by modus tollens when an experiment outcome refutes an hypothesis. More surprising are the results for abductive hypotheses. By exploiting the fact that abductive hypotheses must *explain* observations rather than just be *consistent* with those observations, we are able to prune the abductive hypothesis space with *non-confirming* experiment outcomes as well as with refuting experiment outcomes. Furthermore, we show that this result also holds for the space of consistency-based hypotheses when we restrict Σ to a closed simple causal theory.

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.

5.1 Consistency-Based Differential Diagnosis

Definition 5.1 (Consistency-Based Hypotheses) *A consistency-based hypothesis for Σ and outcome α of the experiment $(true, o)$ is any $H \in conj(\mathcal{H})$ such that $\Sigma \wedge H \wedge \alpha$ is satisfiable.*

Theorem 5.1 (Consistency-Based Differential Diagnosis) *Suppose HYP is the set of all consistency-based hypotheses for Σ, and let α be the outcome of the experiment $(true, o)$. Then*

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

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 $\Sigma \wedge \alpha \wedge H$

[3] Notice that the theorem is stated only for simple experiments 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 full paper.

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 $H \in NEWHYP$.

5.2 Abductive Differential Diagnosis

Contrary to intuition, the criterion for rejecting abductive hypotheses is not simply refutation as demonstrated in the following example.

Example

Recall the definition of an abductive hypothesis: H (a conjunction of literals drawn from some distinguished vocabulary) is an *abductive hypothesis* for the observation o iff $\Sigma \wedge H \models o$ and $\Sigma \wedge H$ is satisfiable. Let Σ be the sentence $h_1 \supset o$, and suppose that the hypotheses are drawn from the vocabulary $\{h_1, h_2\}$. Finally, suppose the initial set 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 o of the experiment $(true, o)$, the set of abductive hypotheses is

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

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

Abduction demands that $\Sigma \wedge H \vDash o$. Hence, by definition, hypotheses that confirm o are abductive hypotheses. However, all other hypotheses that are consistent with Σ but for which $\Sigma \wedge H \not\vDash o$, are not abductive hypotheses. Thus, an experiment outcome that does not confirm an hypothesis, whether it explicitly refutes it or not, causes that hypothesis to be rejected. This is stated formally in the following theorem.

Definition 5.2 (Abductive Hypotheses) *An abductive hypothesis for Σ and outcome α of the experiment $(true, o)$ is any $H \in conj(\mathcal{H})$ such that $\Sigma \wedge H \models \alpha$ and $\Sigma \wedge H$ is satisfiable.*

Theorem 5.2 (Abductive Differential Diagnosis) *Suppose HYP is the set of all abductive hypotheses for Σ, and let α be the outcome of the experiment $(true, o)$. Then*

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

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 results for consistency-based hypotheses.

5.3 Consistency-based Differential Diagnosis of Causal Theories

Poole [11] and Konolige [6] 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 [2] 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 5.3 (Simple Causal Theory [6]) *Let $\mathcal{L}$ be a propositional language. A simple causal theory is a tuple (C, E, Σ) where*

- *C, a set of atomic sentences of $\mathcal{L}$, is the set of causes.*
- *E, a set of atomic sentences of $\mathcal{L}$, is the set of effects we might observe and whose causes we seek.*
- *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.*

Definition 5.4 (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 [2] of Σ.*

The above definition follows from Theorem 1 in [6].

Theorem 5.3 (Consistency-based Differential Diagnosis of Σ^*) *Suppose that $(\mathcal{H}, E, \Sigma)$ 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 experiment $(true, o)$, where $o \in E$. Then*

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

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 $\Sigma^* \wedge H$ 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 [6] 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 experiment outcomes to be generated deductively.

6 Future Work

This paper is a small step towards a theory of experiments for hypothetical reasoning. A variety of approaches and problems remain to be addressed, among which are the following:

1. Probabilistic information (de Kleer and Williams [4]) or other utility measures such as cost or speed of an experiment should be taken into account when developing strategies for ordering experiments. Probabilistic information may also be used as a measure of belief.

2. Extending the results of section 5 to the case $A \neq true$.

3. One issue which seems barely to have been explored is the role of planning in experimental design (Webber et al. [18]). 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. The other is to achieve a suitable *state of knowledge* on the part of the experimenter, for example by taking a person's temperature in order to *know whether* she has a fever. These are quite different. Both may be modeled in the situation calculus (McCarthy and Hayes [8]), but the latter requires formalization in an epistemic logic, along the lines of (Moore [9]). We are currently investigating the use of situation calculus planning formalisms for these and other related problem in hypothetical reasoning.

4. The role of experiments in diagnostic problem-solving other than differential diagnosis.

References

[1] 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.

[2] 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.

[3] J. de Kleer, A.K. Mackworth, and R. Reiter. Characterizing diagnoses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 324–330, 1990.

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

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

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

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

[8] 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.

[9] 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.

[10] D. Oldroyd. *The Arch of Knowledge*. Methuen, 1986.

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.