

Information Service

Vuorimiehentie 5

SF-02150 ESPOO, FINLAND

Telex 125175, Telefax + 358 0 455 4073

University of Waterloo
to Publ. sales

20.dez. 1989

Waterloo
L Ontario N2L 3G1
Canada

Enclosed please find two technical reports by Prof. Poole.
The other two you mention I do not have copies and have forwarded
your request to Prof. Poole at the University of British Columbia.
There is no charge for the reports enclosed.

Please send us
☒ publication Susan DeAngelis
☐ photocopy Research Report Secretary
☐ University of Waterloo
Computer Science Dept.
WATERLOO, Ont. N2L 3G1 CANADA

PLEASE ALWAYS REFER TO OUR ORDER NUMBER

ORDERNR:TINF92920 INF/ Salminen
AUTHOR: Poole, D.L.
TITLE: Building Consistent Theories

JOURNAL:

PUBL: Dep. of Computer Science, Univ. of Waterloo, Ontario,

EDITION:

ISBN:

PUBL. YEAR: 1986

NOTES:

COPIES: 1

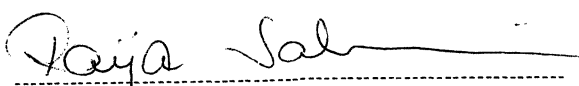
PAYMENT

- ☒ Invoice
☐ Payment enclosed
☐ Deposit account
☐ Charge to credit card
Expires:
☐ Membership

DELIVERY

- ☒ Airmail
☐ Seamail
☐ Special Delivery

TECHNICAL RESEARCH CENTRE OF FINLAND
Information service


Signature

TECHNICAL RESEARCH CENTRE OF FINLAND

ORDER

Information Service

Vuorimiehentie 5

SF-02150 ESPOO, FINLAND

Telex 125175, Telefax + 358 0 455 4073

JAN 4 1990

JAN 05 1990

University of Waterloo
to Publ. sales

20.dez. 1989

Waterloo
Ontario N2L 3G1
Canada

Please send us

- ☒ publication
☐ photocopy
☐

PLEASE ALWAYS REFER TO OUR ORDER NUMBER

ORDERNR:TINF92922 INF/ Salminen

AUTHOR: Poole, D.L.

TITLE: Defaults and Conjectures: Hypothetical reasoning for explanation and prediction

JOURNAL: Technical report CS-87-54

PUBL: University of Waterloo, Canada

EDITION:

ISBN:

PUBL. YEAR: 1987

NOTES:

COPIES: 1

PAYMENT

- ☒ Invoice
☐ Payment enclosed
☐ Deposit account
☐ Charge to credit card
Expires:
☐ Membership

DELIVERY

- ☒ Airmail
☐ Seamail
☐ Special Delivery

TECHNICAL RESEARCH CENTRE OF FINLAND

Information service

Raja Salminen
Signature

TECHNICAL RESEARCH CENTRE OF FINLAND

ORDER

Information Service

Vuorimiehentie 5

SF-02150 ESPOO, FINLAND

Telex 125175, Telefax + 358 0 455 4073

University of Waterloo
to Publ. sales

20.dez. 1989

Waterloo
Ontario N2L 3G1

Canada

Please send us

☒ publication

☐ photocopy

☐

PLEASE ALWAYS REFER TO OUR ORDER NUMBER

ORDERNR:TINF92921 INF/ Salminen

AUTHOR: Poole, D.L.

TITLE: Negation as failure and defaults: a critical comparison

JOURNAL:

PUBL: Dep. of Computer Science, Univ. of Waterloo, Canada

EDITION:

ISBN:

PUBL. YEAR: 1986

NOTES:

COPIES: 1

PAYMENT

☒ Invoice

☐ Payment enclosed

☐ Deposit account

☐ Charge to credit card

Expires:

☐ Membership

DELIVERY

☒ Airmail

☐ Seamail

☐ Special Delivery

TECHNICAL RESEARCH CENTRE OF FINLAND

Information service

Raija Salminen
Signature

Printing Requisition / Graphic Services

15057

1. Please complete unshaded areas on form as applicable.
2. Distribute copies as follows: White and Yellow to Graphic Services. Retain Pink Copies for your records.
3. On completion of order the Yellow copy will be returned with the printed material
4. Please direct enquiries, quoting requisition number and account number, to extension 3451.

TITLE OR DESCRIPTION CS-87-54		DATE REQUIRED ASAP		ACCOUNT NO. 1126626041	
DATE REQUISITIONED July 26/88		PHONE 2192		SIGNING AUTHORITY <i>[Signature]</i>	
REQUISITIONER - PRINT		DEPT.		BLDG. & ROOM NO.	
MAILING INFO -		NAME		<input checked="" type="checkbox"/> DELIVER <input type="checkbox"/> PICK-UP	

Copyright: I hereby agree to assume all responsibility and liability for any infringement of copyrights and/or patent rights which may arise from the processing of, and reproduction of, any of the materials herein requested. I further agree to indemnify and hold blameless the University of Waterloo from any liability which may arise from said processing or reproducing. I also acknowledge that materials processed as a result of this requisition are for educational use only.

NUMBER OF PAGES 50 NUMBER OF COPIES 50		NEGATIVES		QUANTITY	OPER. NO.	TIME	LABOUR CODE
TYPE OF PAPER STOCK <input checked="" type="checkbox"/> BOND <input type="checkbox"/> NCR <input type="checkbox"/> PT. <input checked="" type="checkbox"/> COVER <input type="checkbox"/> BRISTOL <input checked="" type="checkbox"/> SUPPLIED <input type="checkbox"/>		F L M					C 0 1
PAPER SIZE <input checked="" type="checkbox"/> 8 1/2 x 11 <input type="checkbox"/> 8 1/2 x 14 <input type="checkbox"/> 11 x 17 <input type="checkbox"/>		F L M					C 0 1
PAPER COLOUR INK <input checked="" type="checkbox"/> WHITE <input type="checkbox"/> <input checked="" type="checkbox"/> BLACK <input type="checkbox"/>		F L M					C 0 1
PRINTING NUMBERING <input type="checkbox"/> 1 SIDE <input checked="" type="checkbox"/> 2 SIDES PGS. FROM TO		F L M					C 0 1
BINDING/FINISHING <input checked="" type="checkbox"/> COLLATING <input checked="" type="checkbox"/> STAPLING <input type="checkbox"/> PUNCHED <input type="checkbox"/> PLASTIC RING		P M T					C 0 1
FOLDING/PADDING CUTTING SIZE		P M T					C 0 1
Special Instructions Math prints & books enclosed.		P M T					C 0 1
COPY CENTRE OPER. NO. BLDG. MACH. NO.		P L T					P 0 1
DESIGN & PASTE-UP OPER. NO. TIME LABOUR CODE		P L T					P 0 1
TYPESETTING QUANTITY		P L T					P 0 1
P A P 0 0 0 0 0		T 0 1					
P A P 0 0 0 0 0		T 0 1					
P A P 0 0 0 0 0		T 0 1					
PROOF P R F		B I N D E R Y					B 0 1
P R F		R N G					B 0 1
P R F		R N G					B 0 1
		M I S 0 0 0 0 0					B 0 1
		OUTSIDE SERVICES					
		COST \$					
		TAXES - PROVINCIAL <input type="checkbox"/> FEDERAL <input type="checkbox"/> GRAPHIC SERV. OCT. 85 482-2					

**Defaults and Conjectures:
Hypothetical Reasoning for Explanation
and Prediction**

D. L. Poole
Logic Programming and Artificial Intelligence Group
Department of Computer Science

Research Report CS-87-54
October 1987

Defaults and Conjectures: Hypothetical Reasoning for Explanation and Prediction

David Poole

Logic Programming and Artificial Intelligence Group,
Department of Computer Science,
University of Waterloo,
Waterloo, Ontario, Canada, N2L3G1
dlpoole@dragon.waterloo.cdn

October 19, 1987

Abstract

This research follows from the idea that the problem of nonmonotonicity is not a problem with logic, but is a problem with how logic is used. In this paper we consider a simple form of theory formation using normal logic. A system, called Theorist, based on theory formation using a fixed set of possible hypotheses has been built and used on a number of domains. In this paper we investigate a number of distinctions that have been found to be important: between predicting whether something is expected to be true versus explaining why it is true; and between conventional defaults (assumptions as a communication convention), normality defaults (assumed for expediency) and conjectures (assumed only if there is evidence). The effects of these distinctions on recognition and prediction problems are presented, along with algorithms, proofs and examples. This is also argued to be the basis for a theory of model based diagnosis (where there are fault models) and as a framework towards solving multiple extension problems.

1 Introduction

One way to do research into Artificial Intelligence is to argue that we need a certain number of tools and to augment these only when we have shown that they are not adequate for some task. In this way we can argue that we need at least the first order predicate calculus if we want to reason about individuals and relations amongst individuals (given that we want to indirectly describe individuals, as well as talk about the conjunction, disjunction and negation of relations) [Hayes77,Moore82,Genesereth87].

Non-monotonicity has often been cited as a problem with using logic as a basis for commonsense reasoning. In [PGA87] it was argued that instead of deduction from our knowledge, reasoning should be better viewed as a process of theory formation. In [Poole87a] it was shown that default reasoning can be viewed in this way by treating defaults as possible hypotheses that can be used in an explanation.

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

So, we can also argue that we want to do some hypothetical reasoning. This research is part of considering the simplest form of hypothetical reasoning, namely where there is a fixed set of possible hypotheses. This is the framework of the Theorist system [PGA87].

This work is done in the spirit of providing a very limited set of tools. Given a number of tools, we investigate how these tools can be used to solve problems. A repertoire of techniques can then be built to determine how to appropriately use these tools. Only when these tools can be shown to be inadequate, or we have very good reasons why they should be augmented do we add to our set of tools. In this manner, the distinctions outlined in this paper were found from experience by using the system, explaining to others how to use the system and in building applications [Poole87b].

2 Distinctions

Example 1 Consider the following example:

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

On the basis of this knowledge, if we observe that Randy has a bent leg, it is reasonable to hypothesise that he has a broken leg and that the broken leg produced the bent leg. If we subsequently ask whether we predict a sore leg, we would say *yes*, as we hypothesise a broken leg which is typically sore. If we were asked whether we predict, on the evidence of a bent leg, that Randy has head ache we would say *no*, there is no reason to assume that he has a brain tumour given no evidence for it.

This simplistic example indicates a distinction between *explaining observations* and *predicting what we expect to be true*. There is also a distinction between *normality assumptions* (which we want to assume given no evidence to the contrary) and *abnormality assumptions* which we want to assume only if we have evidence.

Each of these distinctions is discussed in this section, and a system which respects such distinctions is outlined in the next section. Formal definitions, outlines of implementations and applications are discussed in later sections.

2.1 Defaults and Conjectures

Example 1 shows a distinction between what I will call *defaults* (or “normality assumptions” which are assumed to be true, given no evidence to the contrary) and *conjectures* (or “abnormality conditions”) which are assumed only if we have evidence (for example, diseases or malfunctions in a system for diagnosis or prototypes in recognition or design tasks).

Defaults and conjectures are similar in that they are both statements that we can hypothesise, but differ markedly in the evidence required to allow the hypothesising.

Defaults can be used unless there is reason to believe otherwise, for example, that some device is working correctly, that if you have broken your leg it is sore, that a bridge will not fall down on top of you. These

can be used to predict or explain observations unless there is evidence that they are incorrect.

This is contrasted with *conjectures* which one has up one's sleeve if one needs to to explain some observation¹. These may include such hypotheses as: someone has some disease, some device is malfunctioning in some way, or there is some object in a scene in a recognition task. Evidence is needed to assume these conjectures.

This distinction seems to be more a notion of difference in degree rather than a difference in kind (for example, in example 1 above, one could say that maybe Randy has a sore head because he may have a brain tumour which would cause a sore head). I would argue that the distinction is important to make, particularly when one must act on one's conclusions (If one doesn't act on one's conclusions, it doesn't seem to matter what one concludes).

I propose that defaults and conjectures should be treated as different categories of possible hypotheses.

There seems to be two alternate ways to incorporate abnormality conditions (those we only want to assume if we have evidence):

1. One alternative is to make the conjectures the conclusions of defaults. Thus if we have conjecture c which we want to consider when a holds, we can have $a \Rightarrow c$ as a default, which is to mean that if we can deduce c from a if it is consistent.

Often the sort of reasoning we are considering is causal reasoning, where for example, if a causes b and we observe b , we want to conjecture a . In Theorist this would be encoded by having $a \Rightarrow b$ as a fact (or a default) and having a as a conjecture. The alternate is to have the evidential default $b \Rightarrow a$ so that when the observation b is added as a fact, we have a as a default conclusion [Pearl87].

There are a number of disadvantages of this alternate approach:

- (a) As pointed out in [Pearl87] there are problems which occur if one does not distinguish between evidential and causal defaults and

¹The user provides the system with formulae that can be used as conjectures if there is evidence for them. The term "conjecture" is used here to mean these formulae that the system has available to conjecture.

their resulting conclusions. Our proposal is an alternate to his; example 7 below shows that the problems that he found with this way of viewing causes does not arise in our proposal.

- (b) As pointed out by McCarthy², it is the mapping from reality to appearance which is much more stable (which corresponds to the causal rules above) than the mapping from appearance to reality (which corresponds to the evidential rules above). It is more reliable to have knowledge of the form “a pen in an image looks like a long thin thing” rather than “a long thin thing in an image is a pen”, because there may be many things which are long and thin. The use of conjectures and defaults allows the use of this more stable mapping.
- (c) Using the evidential rule approach, there is no notion of how causes are grouped. For example, consider a causes both d and e , b causes d and c causes e , and we observe $d \wedge e$. With the conjecture approach, there are two simplest explanations, namely the one with a and the one with b and c . If these were written as evidential rules ($d \wedge e \Rightarrow a$, $d \Rightarrow b$, $e \Rightarrow c$) we can conclude the conjunction $a \wedge b \wedge c$. In this representation there is no notion of the grouping of these conjectures. It also seems strange that we have evidence for the conjunction of all three and not just the disjunction of the three or some other grouping.
- (d) A problem related to the previous occurs when some of the conclusions of evidential rules are inconsistent. For example, if we have a , b and c each causing g , we need the evidential rules $g \Rightarrow a$, $g \Rightarrow b$, $g \Rightarrow c$. If a and b are inconsistent, this gives two extensions; one with a and c and one with b and c . This is a strange result considering the symmetry of the problem.
- (e) there may be no observation which is sufficient in all cases to allow one to hypothesise some conjecture. For example if all one knew was that someone was sneezing, this may be enough to hypothesise that the person has hay fever. If, however we knew a lot of information about the person then knowing that

²AAAI Presidential Address, Austin Texas, 1984.

they are sneezing may not be enough information to allow the hypothesising of hay fever (as some other disease may be a better fit to all of the evidence), even if it was not inconsistent.

2. A second alternative is to only have defaults and to make conjectures the negation of normality defaults (as in [Reiter87]). In this case we assume the negation of a conjecture if it is consistent, that is a conjecture is only concluded if its negation is inconsistent with all other assumptions. There are a number of reasons for preferring not to do this:
 - (a) In a diagnostic setting one may not want to believe that some person does not have a disease, but would rather just not believe that the person does have that disease.
 - (b) It seems to be an unintuitive view of recognition to have the default that a person and a chair and an emu is not at every place in a picture, and to recognise them by finding that these assumptions are inconsistent. Allowing one to hypothesise objects if there is evidence for them seems to be more intuitive.
 - (c) Showing that the negation of some condition is inconsistent is the same as proving the condition. Often there is nothing which allows one to prove that a person is in the image or that someone has some disease (albeit conditioned on some other assumptions). It is better to say that having some disease accounts for the evidence.
 - (d) There may be lots of diseases which cause some normality condition being wrong. There is a difference between saying that there is something wrong with some component and hypothesising what that problem is.
 - (e) In less understood domains than, say, circuit diagnosis, there will often be no symptoms which are actually inconsistent with some diseases or their absence. Diseases interact and one often wants possible grouping of diseases. The system outlined in this paper allows one to build a model of how diseases interact.
 - (f) If we have shown the negation is inconsistent, then we have proven the conjecture based on other assumptions. Thus, the

sort of knowledge required is of the form of evidential rules rather than just the causal rules. All of the disadvantages pointed out before then arise.

The solution proposed here of distinguishing between defaults and conjectures does not get into these problems.

2.2 Normality defaults and conventional defaults

We can distinguish two types of defaults:

- reasonable assumptions, which may be incorrect, but for the time being we will assume that they are true.
- communication conventions, where we know that something is true if we have no statement to the contrary.

An example of the second is the following

We send out invitations to a party saying that we assume that single people are coming by themselves and attached people are coming with their partner, unless they tell us to the contrary. If Bruce is single and accepts the invitation without mentioning a partner, then according to our convention he is not bringing someone else. If he tells us that he is bringing someone else, then that is OK, as our convention was only a default. If we know Eric is married and he doesn't mention that he is coming by himself then we know he is coming with his wife. If Bruce brings someone else, or Eric doesn't bring his wife, it is reasonable to get mad at them because they mislead us.

This seems to use a different sort of default to having something typically being true, and so assuming it for expediency. For example, if we know that broken legs are typically sore, it is reasonable to predict that someone with a broken leg has a sore leg; they may not, however, but I would still not plan a long hike for their visit. We need to make some guesses to get anywhere.

The classic AI example is that we have the default that birds fly, and know that Tweety is a bird, and know nothing else about tweety, we conclude that Tweety flies. How one interprets this conclusion depends on whether the default is a normality default or a conventional default. If it is the former, the answer should mean that “we expect that Tweety flies, as birds typically do, but maybe she doesn’t”; if the default was a conventional default, the answer should mean that “Tweety flies, as if she didn’t fly you would have told us according to the convention that we have between us”. I would claim that the second is still using default reasoning, but this distinction seems to be the distinction that Moore [Moore85] was making when he claimed that autoepistemic reasoning was not default reasoning.

This distinction is also important in solving the “multiple extension problem”. Multiple extensions seem natural and to be expected for normality conditions, where if some individual is in two classes which normally have incompatible properties, it is to be expected that we can predict different things based on the two classes. For conventional defaults, multiple extensions indicate a bug in our convention, as we have evidence that there is a consistent conclusion which we can draw which is incorrect (one of the extensions must be incorrect, as they all can’t be correct as multiple extensions are always incompatible).

These defaults, however, seem to be *used* in the same way (this is supported by [Konolige87], where the formal equivalence between Default logic [Reiter80] and Autoepistemic Logic [Moore85] was proved). For the rest of this paper we will put both of these into one class called the *defaults*.

2.3 Prediction and Explaining Observations

In example 1 we saw a distinction between predicting what we expected to be true versus explaining actual observations made about some system.

There are a number of reasons for such a distinction:

1. there are some things which we only want to hypothesise if we have evidence. We don’t want to hypothesise an invisible person in a picture or a rare disease in a patient if there is no evidence for them. There are different things we bring to bear when asked whether we expect something is true or whether we are told that something is true and asked to find a plausible explanation of why it is true.

2. If we are trying to explain the observation g , it seems irrelevant that $\neg g$ is also explainable; this just means that in some other circumstances g is not true. If, however we are asked whether we predict g , it seems very relevant whether we can also explain its negation.
3. an observation is like a fact, in the sense that all of our theories must be consistent with it (in fact, in the proposed system they imply the observations) whereas a prediction may or may not be explained or consistent with all future theories.

For an explanation of an observation we want to be able to hypothesise both defaults and conjectures; one useful heuristic is that we want each scenario with minimal conjectures as our possible explanations.

When making predictions, we want to be only able to hypothesise defaults (and not conjectures for which we have no evidence). It is argued in section 3.1 that we want to predict things which are in all extensions.

This distinction between prediction and explaining observations is a difference in kind, not a difference in degree (this is important to avoid the question *why isn't there a continuum of values between them?*).

2.4 Facts and Constraints

The other part of the formalism is the class of *facts*. These are intended to be things which are true in the domain we are considering. More precisely they are things that we are not prepared to give up for the sake of the computation.

It is assumed for the sake of the computations that the facts are consistent³. This is a useful assumption to make when implementing systems, as we do not want to have to worry about checking inconsistencies in the facts themselves.

One thing that happens to useful is the ability to prune the set of scenarios without adding new facts. For example, we may want to say that default δ is not applicable under condition c , without always being able to explain $\neg c$ by assuming δ . For example, if we know someone is not guilty,

³Formally we do not need to make this assumption, as if the facts are inconsistent (using classical logic), nothing is explainable but everything is provable.

they should not be a suspect, however we should not conclude that they are guilty just because they are a suspect. This leads us to the class of *constraints*⁴ [Poole87a,Gagné87]. Constraints are formulae which must be consistent with any scenario, but are not part of that scenario. If we add $\neg(c \wedge \delta)$ to the constraints, this prevents both c and δ being in a scenario, without allowing the explanation of $\neg c$ by assuming δ . There seems to be no elegant way to do this without inventing the category of constraints. Constraints can also be used to prevent the contrapositive of a default being used [Poole87a].

2.5 Facts and Defaults

One of the questions that arises when one is using Theorist is when should some piece of knowledge be a fact and when should it be a default. The answer is that it is relative to the problem at hand. One person's facts may be another's hypotheses. This should not be seen as a bug in the theory, but as a feature.

The facts are those pieces of knowledge that for the sake of some argument we are not prepared to give up. A default is some piece of knowledge we are prepared to give up if there is evidence to the contrary. In a similar way that our answers will be conditioned on the defaults used to conclude them, the set of explanations is conditioned by the meta-level assumptions made in building the knowledge base (these may or may not be explicit). Assumptions are made when building a knowledge base; if these are found to be wrong, we try to debug the knowledge base. This framework is the same theory formation and revision framework that the reasoning system itself uses.

One may often want to condition diagnoses with “assuming that the diseases are not acting pathologically and the problem is amongst the known diseases, the diagnosis is ...”. If the symptoms cannot be explained, we know that this assumption is incorrect, and we can try to make explicit our assumptions to try to find out the correct diagnosis. This building of a new layer of the Theorist framework is not any different to the other tasks.

⁴The use of constraints is not essential to the point of this paper, but is included here for completeness and compatibility with [Poole87a]. We tend to be schizophrenic about whether we like them or not.

In the rest of this paper, we assume that we are operating in one level of this hierarchy. The problem of having multiple layers is not considered.

2.6 Facts and Observations

Perhaps a more difficult question is what knowledge should be added as facts and what should be added as observations. Facts and observations are both true of the domain under consideration, but they play very different roles as part of the framework. The answer “observations are those things we observe that need explaining” is a rather vacuous and unsatisfactory answer if there is no way to say what needs explaining and what does not.

Instead I propose a convention that the facts consist of the general knowledge about the domain as well as physical constraints that we are not prepared to give up. The observations are all the things we observe about the particular case in hand. As far as the user is concerned, all she sees about a particular case are observations. The designer of the system can decide that some observations can be treated as facts by making those possible observations as conjectures (see section 5.3.2). This is perfectly consistent with the idea that conjectures are the base causes that we can hypothesise if we have evidence.

2.7 What this is not

This paper is not intended to be a theory of how one changes one's beliefs (i.e., how one changes from attending one scenario to another). That seems to be either the role of a psychological theory (e.g., *How many scenarios do people consider at once? How many scenarios do people consider at all? How much evidence is required before someone changes their mind?* [Harman86]) or an implementation decision (e.g., *Should we build one theory at a time and undo relevant assumptions if we get into trouble?* [Doyle79] or *should we try to build all theories at once?* [de Kleer86]). Both of these are very important issues but are not the subject of this paper.

This is intended to be a competence theory and not a performance theory of nonmonotonic reasoning. This paper talks about consistency as something which can and should be checked in order to hypothesise something. It does not consider that people jump to conclusions with very

little reasoning and only fix up their beliefs when they are convinced they are inconsistent, nor does it talk about how the processes can be done in real time. The psychological validity of this theory is not what is being considered here, neither are very efficient proof procedures.

Although this is presented in a theory formation framework, the proposed system is not intended to be a learning system. There is no way in this framework to generate new hypotheses. We are not trying to automatically generate general theories which are applicable to other cases.

3 Formal Semantics

We assume that we are given a standard first order language over a countable alphabet [Enderton72]. By a formula we mean a well formed formula in this language. By an instance of a formula we mean a substitution of terms in this language for free variables in the formula.

The semantics is defined in terms of three sets of formulae. For different purposes what we are considering to be given and what we take as our possible hypotheses may vary.

A a set of closed formulae which we are taking as given (usually these are the facts, or some current theory of the world under consideration),

C a set of closed formulae taken as constraints⁵, and

H a set of formulae which we take as the possible hypotheses; the elements of *H* are the generators of the formulae that are allowed to be hypothesised. For some purposes these will be just the defaults and for other purposes *H* will be the defaults and the conjectures.

Definition 1 *a scenario of A, H is a set $D \cup A$ where D is a set of ground instances of elements of H such that $D \cup A \cup C$ is consistent.*

This definition is intended to be independent of the logic being used. We want a scenario to be possibly true in the world under consideration (whether it is the real world or some imaginary world), so at least it should

⁵Usually we have *C* as implicit in the following definitions; we will not mention it explicitly, but will assume that scenarios are consistent with the constraints.

be consistent. In this paper the first order predicate calculus is used as the logic for Theorist.

Definition 2 *If g is a closed formula then an **explanation** of g from A, H is a scenario of A, H which implies g .*

That is, g is explainable from A, H if there is a set D of ground instances of elements of H such that

$$\begin{aligned} A \cup D &\models g \text{ and} \\ A \cup D \cup C &\text{ is consistent} \end{aligned}$$

$A \cup D$ is an explanation of g .

Definition 3 *an **extension** of A, H is the set of logical consequences of a maximal (with respect to set inclusion) scenario of A, H .*

In [Poole87a] the correspondence between this definition of extensions and the definition of [Reiter80] (where $\delta \in H$ corresponds to the default : δ/δ in [Reiter80]) was proved. The following theorem was proved in [Poole87a] and follows from the compactness theorem of the first order predicate calculus [Enderton72]

Theorem 1 *g is explainable from A, H iff g is in some extension of A, H*

3.1 Prediction

When predicting what we expect to be true, the possible hypotheses we are prepared to use are the set Δ of defaults. The given A is either the facts or some other scenario about the world. We want to predict some proposition g based on A and Δ if, assuming that everything that is not known to be acting “abnormally” is acting “normally”, we expect that g is true.

Definition 4 *We **predict** g based on A, Δ if g is in every extension of A, Δ .*

If g is not in every extension of A, Δ , there is some scenario S of A, Δ , such that g is not explainable from S, Δ (see theorem 5, below, for a proof of this). Based on our normality conditions and what we are given we cannot rule out S , and so we should not predict g .

Note that we do not predict something if we can just explain it, as we may be able to explain it and its negation. It seems wrong to both predict some proposition and also predict its negation. It is also not adequate to predict some proposition because we can explain it and cannot explain its negation. Consider an example where we can explain a and can also explain $\neg a$. We do not want to predict that a is true. If the only rule about g is $a \Rightarrow g$, then if we can't predict a , we do not want to predict g , even though there is no way to explain $\neg g$.

There seems to be two other candidate definitions for prediction

1. To predict only what is logically implied by an explanation of our observations. This is a very weak notion of prediction which does not allow the use of defaults. Note that if $A \models g$ then g is in every extension of A , so that the above definition is more liberal in its predictions than this alternative. Note that this alternate definition of prediction has one nice property, namely that if A predicts g , and we subsequently observe $\neg g$, then we know that A is wrong. In the definition above, if we subsequently observe $\neg g$, we have to update A , but cannot just reject it.
2. To predict what is in one extension [Reiter80]. That is, to predict g if g is explainable. This is using prediction in the sense of g is predicted if g is "maybe" expected to be true. The difference between this case and predicting only what is in all extensions arises only arises when we can explain some proposition and also explain its negation (if we can't there is only one extension, and the definitions coincide). When predicting what is in all extensions, we are just eliminating from the set of predictions those propositions for which they and their negation can be explained, and any propositions which can be explained only by virtue of assuming one of those is true. Note that according to definition 4, if either a or b are in every extension then $a \vee b$ is in every extension. Although scenarios are conjunctions of formulae, what is predicted is the disjunction of each extension.

3. An alternative is to have a probabilistic prediction. In this case we can assign a weight (e.g., probability or some utility value) to each extension (or to each scenario), and have the probability of some proposition as the weighted sum of all of the extensions in which it is in. Predicting what is in every extension, is then the case of what we can predict with probability 1, assuming that the extensions are covering (i.e., if we assume that the extensions cover all possibilities, then what is in all extensions, is what must be true). This idea is not pursued more in this paper, but is covered in more detail in [Neufeld87b]. (i.e., if we have enough

So without any preference criteria for scenarios (see section 4), it seems as though definition 4 is the correct definition for prediction.

3.2 Explaining Observations

When we are explaining actual observations, we want to build a scenario as to *why* those observations could have occurred. We want to be able to hypothesise conjectures and defaults which would account for these observations.

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

$$F \cup P \cup D \models O \text{ and} \\ F \cup P \cup D \cup C \text{ is consistent}$$

The pair $\langle P, D \rangle$ is said to be the assumptions of the explanation.

4 Scenario Comparators

As first noticed by William of Occam in the twelfth century, not all explanations are born equal.

In this section we want to describe features that we believe such comparators should possess and show on what axes we expect them to exist⁶.

There are a number of axes in which we can compare scenarios. The first is the use to which the scenario is being put. There can be comparators for explanations of observations as well as comparators for making predictions.

For explaining observations we would expect comparators to say when one explanation of the observations is better (for some use) than another.

For prediction comparators, we would expect that some extensions will be preferred over others. Instead of predicting what is in all extensions we predict what is true in all preferred extensions. An example of this are preferring the most specific default [Poole85] when there is more specific and more general knowledge we prefer to use the most specific knowledge we have available. Another example is in preferring the chronologically most persistent extension [Goebel87, Goodwin87] in temporal reasoning when using frame defaults.

An extension can be seen as the complete world description based on a maximal set of things acting normally. As such, we could define preference over extensions and predict g if g is in all preferred extensions. Alternatively, we would expect to predict g if all of the scenarios which explain g are preferred over those scenarios from which g cannot be explained.

The other distinction to make is between

heuristic comparators where all we are saying is that one scenario is preferred over another, but it could be the case that a non-preferred one is the correct one. An example of this is the use of probability to discriminate amongst diagnoses in a diagnostic system [Neufeld87a]. In these cases the predictions should be weighted according to the likelihood or utility of the scenarios which support or reject such predictions. We should take all scenarios into account.

semantic comparators where a less preferred scenario is known to be not the scenario that is wanted. An example of this is to have a preference for more specific defaults in a communication convention. Here the

⁶Much of this section formalises distinctions given in [Neufeld87a]. Here we concentrate on the semantics comparators rather than the heuristic comparators that were discussed in that paper.

extension with the more general default is just wrong. For example we could say that birds typically fly and emus typically don't fly and that Polly is an emu and if we have the convention of using the most specific default then we know that Polly does not fly as we have more specific knowledge about emus than the more general knowledge about birds. These semantic comparators are used to prune the space of acceptable scenarios. They are called semantic because they are making semantic statements about some scenarios not being correct.

Note that prediction comparators are used when explaining observations. For example, if we have the semantic prediction preference for the most specific default, and have *birds fly*, *emus don't fly*, and *flying emus fly* as defaults, and observe that *polly* is an emu and flies, then we don't want to say that *polly* is a bird and so flies and *polly* is an *emu*, its just that the default that *polly* does not fly cannot be used as it is inconsistent with what is known. We have to say that emus not flying is more specific than birds flying means that the default about birds flying is not applicable to *polly*. The desired answer is that *polly* is a special type of emu, namely a flying emu.

4.1 Three Comparators for Explaining Observations

Three useful comparators for explanations are discussed here:

1. the *minimal explanation*, that is to prefer the explanations that makes the fewest (in terms of set inclusion) assumptions⁷.
2. the *least presumptive explanation*. Explanation E_1 is less presumptive than E_2 if $E_2 \models E_1$. That is, if E_1 makes less (in terms of what can be implied) assumptions than E_2 .
3. the *minimal abnormality explanation*. Explanation E_1 with conjecture assumptions P_1 and default assumptions D_1 is less abnormal than E_2 with assumptions $\langle P_2, D_2 \rangle$ if $E_2 \models P_1$ and either $E_1 \not\models P_2$

⁷Note that I am not advocating comparing scenarios by counting the number of assumptions in them. Such comparators have too many problems of slight rerepresentations of the problem domain giving different answers.

or $(E_1 \models P_2 \text{ and } E_2 \models D_1)$. That is, if it makes less abnormality assumptions or it makes the same abnormality assumptions and fewer normality assumptions.

The first two can both be seen as preferring the minimal explanation, the first if we treat a scenario as a set of axioms, the second if one equates a scenario with its logical theory (i.e., considers the set of all logical consequences of the explanation). The first two comparators can be seen as a *semantic* comparators in that if there is one correct⁸ explanation, there is a minimal and a least presumptive explanation which is also correct.

I cannot think of a situation where one would not want the minimal explanation (i.e., why one would want to make extra unneeded assumptions). Although there are cases where no least presumptive explanation exists (see example 4 below) as well as cases where it can be argued that the least presumptive explanation may not be the “best” explanation (see example 5 below), it seems as though the least presumptive explanation is often the desired explanation.

Example 2 Let

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

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

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

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

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

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

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

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

⁸Correct in the sense of true in the intended interpretation. This is a property which can be verified by an oracle observer of the system.

The third definition is more heuristic. It may be the case that the minimal abnormality explanation is not the correct one, however we may not want to consider more abnormalities than we have evidence for.

The following two theorems give the relationships between these three comparators.

Theorem 2 *A least presumptive explanation is always equivalent to a minimal explanation.*

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

Note that this does not mean that a least presumptive explanation is always a minimal explanation, as we can always throw in hypotheses and conjectures implied by a least presumptive explanation into the explanation and it is still least presumptive, but no longer minimal.

Theorem 3 *A minimal abnormality explanation is always a least presumptive explanation.*

Proof: Suppose E is a minimal abnormality explanation with assumptions $\langle P, D \rangle$. We need to prove that there cannot be an explanation which is strictly less presumptive than E . Assume that explanation E' , with assumptions $\langle P', D' \rangle$, is strictly less presumptive than E (i.e., $E \models E'$ and $E' \not\models E$); we want to show that E' is strictly less abnormal than E .

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

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

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

The converse is not always true.

Example 3 Consider the following system

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

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

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

There are other explanations for the observation, for example

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

but all of these make extra assumptions for which we have no evidence (and, together with F , imply the least presumptive explanation).

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

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

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

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

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

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

One problem that arises is that there may not be a least presumptive explanation:

Example 4 Consider the following system:

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

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

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

Example 5 Suppose we are building a user modelling system, and want to be able to conjecture the interests of people and have the following conjectures:

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

The defaults of the interests are given as defaults:

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

If we observe that someone borrows logic books then it is reasonable to conjecture that they are interested in logic. This is the least presumptive explanation. If we observe that someone borrows logic books and writes computer programs then there is one least presumptive explanation, namely that they are interested in computer science and interested in logic. The alternate explanations, namely that they are interested in formal AI or interested in hardware are not going to be least presumptive, although one could argue that they are the best explanations. The problem here is that there is some notion of simplicity in gauging the best explanation. It is best to get to the root cause of the problem rather than just giving the weakest explanation.

This is similar to what was argued in [Popper62, p. 219] that one does not always want the most likely explanation. He proposes a *verisimilitude* for comparing theories. [Quine78, chapter 6] defined five *virtues* on which to compare explanations.

Much more work needs to be done on defining appropriate scenario comparators.

5 A Default Reasoning System

5.1 Architecture

The architecture we are considering is one where the system is provided (see section 5.2 for how this is done) with the facts, constraints, defaults and conjectures. We assume that these provide the general knowledge about the domain that is being modelled (i.e., how diseases interact and how symptoms work in a diagnosis system, and general knowledge about objects, occlusion etc., in a recognition task), all specific knowledge about the particular case is added as observations (see section 5.3 for a description of the programming methodology).

We have a sequence of observations given to the system which builds up its best (according to the explanation comparisons given) explanations of the observations. From each of these explanations we can ask what they predict and what is expected. The system can also propose what observations it would like about the world in order to prune and refine its explanations.

5.2 Interacting with the system

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

The state of the system can be described as a tuple

$$\langle F, C, \Delta, \Pi, O, \mathcal{E} \rangle$$

where

F is the set of facts

C is the set of constraints

Δ is the set of defaults

Π is the set of conjectures

O is the set of observations that have been made

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

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

fact w .

where w is a formula, means “ $\forall w$ ”⁹ is a new fact. That is the resulting state is $\langle F \cup \{\forall w\}, C, \Delta, \Pi, O, \mathcal{E}' \rangle$ where \mathcal{E}' is the resulting

⁹ $\forall w$ is the universal closure of w , that is, if w has free variables \bar{v} then $\forall w$ means $\forall \bar{v} w$. Similarly $\exists w$ is the existential closure of w .

explanations given $\forall w$ as a fact (see section 6.1 for a description of how this can be computed).

constraint w .

where w is a formula, means “ $\forall w$ ” is a new constraint. That is, the new state is $\langle F, C \cup \{\forall w\}, \Delta, \Pi, O, \mathcal{E}' \rangle$, where \mathcal{E}' is the set of new minimal explanations.

default n .

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

$$\langle F, C, \Delta \cup \{n\}, \Pi, O, \mathcal{E}' \rangle$$

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

default $n : w$.

where w is a formula, and n is a name means that w is a default, with name n . Formally this means that the new state is

$$\langle F \cup \{\forall(n \Rightarrow w)\}, C, \Delta \cup \{n\}, \Pi, O, \mathcal{E}' \rangle$$

See [Poole87a] for a discussion on naming defaults.

conjecture n .

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

$$\langle F, C, \Delta, \Pi \cup \{n\}, O, \mathcal{E}' \rangle$$

conjecture $n : w$.

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

$$\langle F \cup \{\forall(n \Rightarrow w)\}, C, \Delta, \Pi \cup \{n\}, O, \mathcal{E}' \rangle$$

observe g .

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

predict g, S .

where g is a formula and S is a scenario (usually one of the elements of E), returns *yes* (together with the instance) if some instance of g is in every extension of S and *no* otherwise (not that we predict that g is false, but rather that we do not predict that it is true).

predict g .

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

It is assumed that essentially all of the “general knowledge” about a system is added as facts, and that all specific knowledge about the particular case at hand is added as observations.

Example 6 Example 1 can be specified as follows:

conjecture *brain-tumour*.

conjecture *broken-leg*.

default *tumoured-heads-ache*: *brain-tumour* \Rightarrow *head-ache*.

default *broken-legs-are-sore*: *broken-leg* \Rightarrow *sore-leg*.

default *broken-legs-are-bent*: *broken-leg* \Rightarrow *bent-leg*.

If we make the observation

observe *bent-leg*.

we have one minimal and least presumptive explanation:

$$\{\textit{broken-leg}, \textit{broken-legs-are-bent}\}$$

If we subsequently ask:

predict *head-ache*.

the answer is *no*. If we ask

predict *sore-leg*.

the answer is *yes*.

Example 7 (Pearl) [Pearl87, p. 371] gives the following example to argue that there should be a distinction between *causal rules* and *evidential rules*. Here we show how the problems he was trying to solve do not arise in our system. We add the causal rules as defaults (or facts if we do not want to consider them having exceptions)

```

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

```

Instead of adding the reverse of these rules as Pearl does, we make the possible causes we are considering as conjectures:

```

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

```

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

$$\{rained-last-night\}$$

From this we can predict that the grass is wet, that the grass is cold and shiny and that my shoes are wet. There is no way to predict that the sprinkler was on last night (which was the problem with having the evidential rules as explicit rules (see section 2.1)).

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

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

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

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

5.3 Programming Methodology

It is not adequate to just define a representational language and leave it at that; it is also necessary to say how this language can be used to solve the sorts of problems we want to solve. This knowledge can only come from

experience with using the system. In this section we discuss some useful ways to use the system that we have found.

Essentially statements that one would expect to be true under normal circumstances should be added as defaults. Cases where one would not want to assume these would be added as facts or constraints (depending on whether one wants to be able to conclude other things from assuming the default). For anything which could possibly be observed, one has to consider what an appropriate explanation for that observation would be. This may be the observation itself (section 5.3.2) or some more complex formula where the observation is broken down into more primitive observations which would in turn need to be explained. The implication of the observation from these causes can be any mixture of facts, defaults and conjectures. So for example if g is a possible observation and c is a possible cause for g , then c and $c \Rightarrow g$ should be considered either as facts, defaults, conjectures or other observations which need in turn to be explained. There is nothing in the formalism which makes us think that $c \Rightarrow g$ should be a default and c a conjecture (although, indeed they may be).

5.3.1 Parameterising Possible Hypotheses

When building systems using Theorist it is important to know how the way possible hypotheses can be parameterised to have different effects.

In general the free variables in possible hypotheses are the values on which the truth of the hypothesis depends. If, for example, the truth of a hypothesis depends on the time, then time should be a parameter of the possible hypothesis (then contradicting it at one time should not contradict it for other times). If the identity of some parameter is irrelevant to the truth of a hypothesis, then it should not be free in the possible hypothesis.

Example 8 Consider the statement “you may assume that a person likes all dogs”. This can be used to predict that some person likes some dog unless there is evidence to the contrary. If there is one dog which they do not like then we cannot assume that they like other dogs. This can be given by

default *likes-all-dogs*(P) : *person*(P) \wedge *dog*(D) \Rightarrow *likes*(P, D).

which means $likes\text{-}all\text{-}dogs(P)$ is an element of Δ and

$$\forall P \forall D \text{ person}(P) \wedge \text{dog}(D) \wedge \text{likes}\text{-}all\text{-}dogs(P) \Rightarrow \text{likes}(P, D)$$

is a fact. By making P a parameter of the default, and not D means that the default is contradicted for a person if there is one dog they do not like. For example, given also

fact $\text{dog}(\text{fido})$.
fact $\text{dog}(\text{honey})$.
fact $\text{person}(\text{randy})$.
fact $\text{person}(\text{sumo})$.
fact $\neg \text{likes}(\text{randy}, \text{fido})$.

we can explain $\text{likes}(\text{sumo}, \text{fido})$ but cannot explain $\text{likes}(\text{randy}, \text{honey})$, as $\text{likes}\text{-}all\text{-}dogs(\text{randy})$ is inconsistent with the facts.

This should be contrasted to the statement “you may assume that any person likes any dog”. Here the existence of one dog that a person does not like should not prevent us from assuming they like other dogs. This can be specified by

default $\text{likes}\text{-}dog(P, D) : \text{person}(P) \wedge \text{dog}(D) \Rightarrow \text{likes}(P, D)$.

which means $\text{likes}\text{-}dog(P, D)$ is an element of Δ and

$$\forall P \forall D \text{ person}(P) \wedge \text{dog}(D) \wedge \text{likes}\text{-}dog(P, D) \Rightarrow \text{likes}(P, D)$$

is a member of F .

In this case, from the above facts we can explain $\text{likes}(\text{sumo}, \text{fido})$ and $\text{likes}(\text{randy}, \text{honey})$ but not $\text{likes}(\text{randy}, \text{fido})$

In contrast to other proposals where the hypotheses must be consistent with our observations (e.g., [Reiter87, de Kleer87]), our hypotheses must have the power to predict the observations. To do this the conjectures should be parameterised by the relevant inputs on which the cause depends as well and the possible outputs.

For example if we want to consider malfunction d , and it depends on parameters I_1, \dots, I_n (for example, incoming current, time of day, temperature in Antarctica) and predicts values for O_1, \dots, O_m (for example, temperature

of a person, output current) then the conjecture should be specified as being parameterised by all of these, namely as $hasmal_d(I_1, \dots, I_n, O_1, \dots, O_m)$. We are then allowed to hypothesise that the system has some outputs for the inputs given as

fact $input(I_1, \dots, I_n) \wedge$
 $hasmal_d(I_1, \dots, I_n, O_1, \dots, O_m) \wedge$
 $reln(I_1, \dots, I_n, O_1, \dots, O_m)$
 $\Rightarrow output(O_1, \dots, O_m).$
constraint $c(I_1, \dots, I_n, O_1, \dots, O_m) \Rightarrow$
 $\neg hasmal_d(I_1, \dots, I_n, O_1, \dots, O_m).$

Where *reln* is some relation that must hold between the inputs and the outputs before we can use the hypothesis to predict the output for the given input, and *c* is some relation which cannot hold between the input and the output. If we observe some output produced from some input, and if it fits the constraints of the malfunction (i.e. *reln* is true of them, and we cannot prove that *c* is true of them) then the appropriate instance of *d* can be conjectured as a cause of the output.

Example 9 Consider the problem of having a lamp connected to a battery. Suppose if the battery is acting normally its voltage is between 1.2 and 1.6; if it is overcharged, above this, and if it is flat the voltage is below this range. The lamp will be normally be lit if the voltage is over 1.3 and will be dim if the voltage is between 1.0 and 1.3, however if the voltage ever gets over 1.8 then the lamp will blow and never be normal again.

The following relations are relevant:

voltage(*V*, *T*) means that at time *T* the voltage across the battery (and also across the lamp) is *V* volts.

battOK(*V*, *T*) means that at time *T*, the battery is working OK and is producing *V* volts.

overcharged(*V*, *T*) means that at time *T*, the battery is overcharged and is producing *V* volts.

flat(*V*, *T*) means that at time *T*, the battery is flat and is producing *V* volts.

lampOK(T) means that at time T , the lamp is working normally.

dim(T) means that the lamp is dim at time T .

lit(T) means that the lamp is lit at time T .

We can specify that the battery normally produces some voltage between 1.2 and 1.6 by

```

fact battOK( $V, T$ )  $\Rightarrow$  voltage( $V, T$ ).
default battOK( $V, T$ ).
fact battOK( $V, T$ )  $\Rightarrow 1.2 \leq V \wedge V \leq 1.6$ .

```

We can also specify how the problems/malfunctions manifest themselves:

```

fact overcharged( $V, T$ )  $\Rightarrow$  voltage( $V, T$ ).
conjecture overcharged( $V, T$ ).
fact overcharged( $V, T$ )  $\Rightarrow V > 1.6$ .
fact flat( $V, T$ )  $\Rightarrow$  voltage( $V, T$ ).
conjecture flat( $V, T$ ).
fact flat( $V, T$ )  $\Rightarrow V < 1.2$ .

```

We can also state that there cannot be two different voltages at any time (Note that this could have also be achieved by making *voltage* a function from time to the voltage at that time.)

```

fact voltage( $V_1, T$ )  $\wedge$  voltage( $V_2, T$ )  $\Rightarrow V_1 = V_2$ .

```

Similarly we can axiomatise how the lamp works normally

```

fact lampOK( $T$ )  $\wedge$  voltage( $V, T$ )  $\wedge V \geq 1.3 \Rightarrow$  lit( $T$ ).
fact lampOK( $T$ )  $\wedge$  voltage( $V, T$ )  $\wedge 1.0 \leq V \wedge V < 1.3 \Rightarrow$  dim( $T$ ).
default lampOK( $T$ ).
fact lampOK( $T$ )  $\wedge$  voltage( $V, T$ )  $\Rightarrow V \leq 1.8$ .
fact  $\neg$ lampOK( $T_0$ )  $\wedge$  before( $T_0, T_1$ )  $\Rightarrow \neg$ lampOK( $T_1$ ).

```

Given no observations, as we would expect, we cannot predict for example that the voltage is 1.5 volts (as that is not true in all extensions), however we can predict

$$\forall T \exists V \ V \geq 1.2 \wedge V \leq 1.6 \wedge \textit{voltage}(V, T)$$

Given no observations, if we were asked to predict whether the lamp is lit at some time t , then the answer is *no*, as $\{battOK(1.25, t)\}$ is a scenario from which $lit(t)$ cannot be explained. We can, however, predict $lit(t) \vee dim(t)$. There are infinitely many explanations of $lit(t) \vee dim(t)$, namely consisting of

$$\{battOK(V, t), lampOK(t)\}$$

for every V such that $1.2 \leq V \leq 1.6$, and there is no scenario of F, Δ from which $lit(t) \vee dim(t)$ cannot be explained.

If we observe $dim(t_0)$, then there are the least presumptive explanations:

$$\{battOK(V, t_0), lampOK(t_0)\}$$

for $1.2 \leq V \leq 1.3$ and

$$\{flat(V, t_0), lampOK(t_0)\}$$

for $1.0 \leq V < 1.2$. Only the first are minimal abnormality explanations. We only predict things which are true in all extensions of these.

If we later observe that the voltage is indeed 1.25 at time t_0 , then there is one explanation, namely

$$\{battOK(1.25, t_0), lampOK(t_0)\}$$

5.3.2 Observations and Facts

In section 2.6 it was claimed that all of the generalised knowledge about a domain should be added as facts and all knowledge about a particular case should be added as observations. This convention makes a clear distinction for the user of the system, but requires the builder of the knowledge base to be aware of this. The conjectures that the designer provides should include all of those possible observations that one want to treat as facts. This is entirely withing the spirit of conjectures; we just don't want a deeper analysis of the cause of these observations.

For example, if we want the age of a patient to be able to be added as an observation, but we do not want a deep analysis of why we think this is the observed age, then we can add

conjecture *age*(*P*, *A*).

If we find out the age of a particular person, then this is added as

observe *age*(*jenn*, 23).

This will have the same effect as adding *age*(*jenn*, 23) as a fact as the least presumptive explanations will always be the ones that contain *age*(*jenn*, 23).

The conjectures should then be whatever we are prepared to accept as explanations for the observations whether they are things which don't really need to be explained or deep causes for complex behaviour.

5.3.3 Causes and Symptoms

One of the ways of looking at recognition and diagnostic tasks is to find the causes of symptoms [Cox87]. There are cases when something can be considered a cause sometimes and symptom at other times. If not handled appropriately, this may become a problem in our system if we prefer the least presumptive explanation (see section 4). This can, however, be fixed by an appropriate structuring of the knowledge base. Consider the following example:

Example 10 Suppose we want to represent the sentences

People sneeze because they have a cold.

Sometimes people just sneeze.

One representation of this may be

conjecture *sneezes*(*X*).

conjecture *has-cold*(*X*).

default *sneezing-because-of-cold*(*X*): *has-cold*(*X*) \Rightarrow *sneezes*(*X*).

If we observe

observe *sneezes*(*eric*).

there is one least presumptive explanation, namely

$\{\textit{sneezes}(\textit{eric})\}$

The explanation that eric has a cold is not considered because it is more presumptive than the other explanation. There is a problem here with interpretation; we should not consider this answer as meaning the second sentence above (i.e. that he is just sneezing for no reason). This answer means that he is sneezing, and that is considered as a cause in itself. It does not exclude that he is sneezing because of a cold.

If, however, we want to distinguish between the two causes then the appropriate way to represent this is

```

conjecture random-irritation(X).
conjecture has-cold(X).
default sneezing-because-of-cold(X): has-cold(X)  $\Rightarrow$  sneezes(X).
default just-sneezing(X): random-irritation(X)  $\Rightarrow$  sneezes(X).

```

In this case there are two least presumptive explanations of eric sneezing:

$$\{ \textit{random-irritation}(\textit{eric}), \textit{just-sneezing}(\textit{eric}) \}$$

$$\{ \textit{has-cold}(\textit{eric}), \textit{sneezing-because-of-cold}(\textit{eric}) \}$$

Here, we have the appropriate diagnoses.

6 Implementation

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

One of the things that we want to know is whether we can do a localised search rather than always having to do a full consistency check. We would like to only search the part of the space that is relevant to what is being added or asked. We do not want to have to search other parts of the space; we would like to know that irrelevant parts of the knowledge base are indeed irrelevant.

One way that this can be done is to only assume a limited form of completeness of the theorem prover. We want our theorem prover to be sound, but only require completeness in the sense that if there is a relevant proof of some goal, it can be found. A proof of g from A (denoted $A \vdash g$) is assumed to be sound (i.e., if $A \vdash g$ then $A \models g$), but it need only be complete

in the sense that if A is consistent and $A \models g$ then $A \vdash g$. Linear Resolution [Chang73] with head clause g is such a proof procedure. Hopefully such deduction systems can be much more efficiently implemented as they do not need to consider irrelevant reasons for something following from a set of axioms.

The following two theorems are important for implementing the system

Theorem 4 *If $A \cup C$ is consistent, g is explainable from A, H if and only if there is a ground proof of g from $A \cup D$ where $D = \{d_1, \dots, d_n\}$ is a set of ground instances of elements of H such that $A \wedge C \wedge \{d_1, \dots, d_{i-1}\} \not\vdash \neg d_i$ for all $i = 1..n$.*

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

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

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

1. For each $d_i \in D$ try to prove g from $A \cup H$, and make D the set of instances of elements of H used in the proof.
2. ground D (make all free variables in D ground). We thus have created a ground proof of g from $A \cup D$.
3. try to prove $\neg d_i$ from $A \wedge \{d_1, \dots, d_{i-1}\}$. If all such proofs fail, D is an explanation for g .

Theorem 5 *The following are equivalent:*

1. g is in every extension of A, H

2. *there does not exist a scenario S of A, H such that g is not explainable from S, H .*
3. *there is a set \mathcal{E} of (finite) explanations of g such that there is no scenario S of A, H such that $\forall E \in \mathcal{E}, S$ is inconsistent with E .*
4. *there is some D such that $A \wedge D \vdash g$ such that if $d \in D$ and $\neg d$ is explainable by D_i , then g is in every extension of $A \wedge D_i, H$.*

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

$3 \Rightarrow 2$. Suppose 3 holds, and there is a scenario S from which g is not explainable. Then each $E \in \mathcal{E}$ is inconsistent with S (otherwise $E \cup S$ is an explanation of g from S, H).

$1 \Rightarrow 3$. Suppose 1 holds. The set of all maximal scenarios has the property given in 3 (except the finite membership). By the compactness theorem of the first order predicate calculus [Enderton72] there is a finite subset \mathcal{E} of the maximal scenarios which imply g . If some S were inconsistent with all elements of \mathcal{E} it would be inconsistent with the maximal scenarios, and we know that such an S cannot exist. So \mathcal{E} is a set which satisfies 3.

$3 \Rightarrow 4$. Suppose 3 holds, then the set \mathcal{E} is countable (as it is a subset of the set of finite strings in a language with finite generators). Let D be the minimum element of \mathcal{E} according to some ordering. Then we know $A \wedge D \vdash g$. As g is in every extension of A, H it is in every extension of $A \wedge D_i, H$.

$4 \Rightarrow 2$. Suppose 4 holds and there is some scenario S such that g is not explainable from S . D is inconsistent with S (otherwise $S \cup D$ is an scenario of S, H which explains g), so there is some $d \in D$ which follows from consistent $S' = S \cup D'$ where $D' \subseteq D$ and so by 4, g is in every extension of S' , and so is in one extension of S' , a contradiction to g not being explainable from S . \square

Note that the set of explanations referred to in 3 is countable, but not necessarily finite. The following example has an infinite set of possible explanations to check.

Example 11 Let $H = \{p(X)\}$

$$F = \{ \begin{aligned} &q(0), \\ &\forall n \, q(n) \Rightarrow q(s(n)), \\ &pos(s(0)), \\ &\forall n \, pos(n) \Rightarrow pos(s(n)), \\ &\forall n \, pos(n) \Rightarrow lt(0, n), \\ &\forall n \forall m \, lt(m, n) \Rightarrow lt(s(m), s(n)), \\ &\forall n \forall m \, lt(m, n) \Rightarrow \neg(p(m) \wedge p(n)), \\ &(\exists x \, p(x) \wedge q(x)) \Rightarrow g \end{aligned} \}$$

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

Point 4 of Theorem 5 leads to the algorithm: to prove that g is in every extension of A, H

1. try to prove g from $A \cup H$, and make D the set of instances of elements of H used in the proof.
2. ground D (make all free variables in D ground). We thus have created a ground proof of g from $A \cup D$. Let $D = \{d_1, \dots, d_n\}$.
3. try to explain $\neg d_i$ from $A \wedge \{d_1, \dots, d_{i-1}\}, H$. If there is an explanation using no assumptions then D is inconsistent; for each other D_i explaining d_i , we try to prove g is true in all extensions of $A \cup D_i$.

The best way of looking at this algorithm is that the third step is trying to construct the scenario which is potentially inconsistent with the part of the other explanations needed to prove g .

6.1 Building and Maintaining the Knowledge Base

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

1. record just what was explicitly told and then compute all answers when asked.
2. maintain one explanation for the observations and build another if this one proves to be wrong (e.g., [Doyle79]).
3. maintaining multiple, but not all explanations. For example, maintaining just those minimal abnormality explanations and only considering others if these prove inadequate. As example 13 below shows, it is often difficult to make sure that one is maintaining the minimal abnormality explanations without also maintaining all of the other least presumptive explanations.
4. maintaining parts of all of the least presumptive explanations. This may make it easier to see when one explanation can be replaced by a better explanation. For example [Neufeld87a] describes an algorithm which always maintains the most likely explanation by maintaining enough of other explanations to ensure that they will be less likely than the preferred one.
5. maintain all least presumptive explanations (or all minimal explanations). The ATMS of [de Kleer86] can be seen as doing this.
6. maintaining a representation of all extensions (e.g., the generating hypotheses). This may make building the knowledge base inefficient, but may make it easier to query.

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

Unless one is not maintaining explanations, we want to know how adding facts, constraints, defaults, hypotheses and observations affects the explanations generated.

6.2 Incremental Observations

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

Theorem 6 *We can build minimal explanations incrementally. That is, if S_1, \dots, S_n are the minimal explanations of g_1 from F, Π, Δ then the minimal explanations of g_2 from the S_i, Π, Δ are exactly the same as the minimal explanations of $g_1 \wedge g_2$ from F, Π, Δ .*

Proof: If E is an explanation of $g_1 \wedge g_2$ from F, Π, Δ then E is an explanation of g_1 from F, Π, Δ , so $\exists S \subseteq E$ such that S is a minimal explanation of g_1 . Then E is an explanation of g_2 from S , and is minimal.

Similarly if E is an explanation of g_2 from some S_i then E is an explanation of $g_1 \wedge g_2$ from F . Hence, if the minimal explanations of g_2 from the S_i are selected then these are the minimal explanations of $g_1 \wedge g_2$ from F . \square

Theorem 7 *If S_1, \dots, S_n are the least presumptive explanations for g_1 from F, Π, Δ then the following are equivalent*

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

Proof: $1 \Rightarrow 2$. Suppose S is a least presumptive explanation of $g_1 \wedge g_2$ from F . Then S is an explanation of g_1 , so one S_i implies S . S is an explanation of g_2 from S_i ; we need to

show that it is least presumptive. Suppose S' is a strictly less presumptive explanation of g_2 from S_i , then it is an explanation of $g_1 \wedge g_2$ from F less presumptive than S , a contradiction to the minimality of S .

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

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

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

Example 12 Let

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

If we observe g_1 there are minimal (and least presumptive) explanations, namely $\{a\}$ and $\{b\}$. If we then observe g_2 , there is one minimal explanation, namely $\{b\}$. Note that we can explain g_2 from $\{a\}$, but this explanation is subsumed by a simpler explanation.

Theorem 7 does not work for minimal abnormality explanations. Consider the following example:

Example 13 Let

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

The least presumptive explanations for g_1 are

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

$$\{a, d_2\}$$

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

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

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

the first of which is the minimal abnormality explanation.

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

In the rest of this section we will assume that the minimal explanations are maintained, and we will consider the effects of making new declarations to the system.

6.3 Adding new facts

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

The command

fact w .

means that the knowledge base is changed from

$$\langle F, C, \Delta, \Pi, O, \mathcal{E} \rangle$$

to

$$\langle F \cup \{\forall w\}, C, \Delta, \Pi, O, \mathcal{E}' \rangle$$

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

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

$$\begin{aligned} F \cup E &\models O \\ F \cup E \cup C &\text{ is consistent.} \end{aligned}$$

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

1. $F \cup E' \models O$ in which case E' is an explanation of O from F . $E' \in \mathcal{E}$ as there can be no smaller explanation of O from F , otherwise it is a smaller explanation of O from $F \cup \{\forall w\}$. We can thus carry over the old explanation over from \mathcal{E} .
2. $F \cup E' \not\models O$ and so $F \cup E' \cup \neg O$ is consistent and implies $\neg \forall w$.

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

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

1. try to explain $\neg \forall w$ from $F \cup \neg O, \Delta \cup \Pi$. The generated explanation should be checked consistent with $F \cup \{\forall w\} \cup C$. We thus only need to consider relevant proofs from the added fact, and not start from scratch building new explanations.
2. try to prove $\neg \forall w$ from $F \cup E \cup C$, for each $E \in \mathcal{E}$, and remove any explanation which is proven inconsistent.

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

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

6.3.1 Adding Constraints

Adding constraints can only remove explanations from the set of explanations by making them inconsistent. We cannot add new explanations, nor can we make one explanation less presumptive when it previously was not. If \mathcal{E} was the set of explanations before the constraint w was added, then $E \in \mathcal{E}'$ where \mathcal{E}' is the set of least presumptive explanations if and only if $E \in \mathcal{E}$ and $F \cup C \cup E \not\models \neg w$. Thus we can just try to prove the negation of the newly added constraint.

6.3.2 Adding Defaults and Conjectures

Consider the problem of adding the default

default $d : w$.

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

Theorem 8 (Semimonotonicity) *If \mathcal{E} is the set of explanations before the default was added and \mathcal{E}' the explanations after then $\mathcal{E} \subseteq \mathcal{E}'$.*

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

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

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

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

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

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

7 A Theory of Diagnosis

In this section I wish to argue that the preceding outline is a good basis for formalising model-based diagnosis. This theory, as a theory of diagnosis, is an attempt to bridge the gap between diagnosis from first principles [Reiter87, Davis84, Genesereth84, de Kleer87], and more experience-based diagnosis based on knowledge as to how diseases and malfunctions normally manifest themselves [Weiss78, Patil81, Popl83, Brown82].

In diagnosis from first principles, one has a model of the intended behaviour of the system. Any discrepancy between the predicted and observed behaviour means that the assumptions that components are working correctly is inconsistent with the observations, and so we can prove that some components are not working correctly. Reiter [Reiter87] defines a diagnosis as a minimal set of assumptions that components are faulty, together with the assumption that all other components are working correctly that is consistent with all observations of the system.

In section 2.1 some arguments were given as to why the Theorist approach is different and has some advantages over Reiter’s methodology. In this section I wish to give a more pragmatic comparison.

When doing a diagnosis, we want to find out what is wrong with some system. We thus want to find out some minimal set of components we need to assume are faulty given our evidence. Somehow we have to make these assumptions relevant to the observations and not to always say that we should just assume nothing. Reiter minimises abnormality assumptions and maximises normality assumptions so that the observations are consistent. In the framework suggested in this paper, we minimise all assumptions, however to stop always degenerating to the case of making no assumptions, we must have our assumptions implying the actual observations.

The main consequences of this distinction is that we have the ability as well as the obligation to state how problems manifest themselves. We must not only state how normal components act, but also how abnormal components act. This is not as big an imposition as it may seem as we can always say that a component is abnormal if it is working in some way that is different to what was designed.

Example 14 (Genesereth and Reiter) This example is derived from [Genesereth84, Fig. 8, p416] and [Reiter87, Example 2.2, p. 60]. To specify the intended action of an and-gate, Reiter give the axiom (here we have modified Reiter's notation slightly to allow multiple observations as in [Genesereth84])

$$andg(X) \wedge \neg ab(X) \Rightarrow out(X, T) = and(in1(X, T), in2(X, T))$$

This axiom tells us what happens if a gate is working normally. It does not tell us what happens if the gate is acting abnormally. By abnormal, Reiter means that there exists some value for which it gives the incorrect value.

In Theorist, we parameterise the assumption so that we can talk about acting normally for some inputs and acting abnormally for other inputs. If we decide that the relevant parameters to the normality assumption are the inputs to the gate (i.e., not on the time of day¹⁰, or the amount of money in my bank account), then we use the relations $ab(X, I_1, I_2, O)$ which means that gate X is working abnormally for inputs I_1 and I_2 , and producing

¹⁰By not making the value depend on the time, we are making the non-intermittency assumption, namely that the value of the outputs of a gate depends only on the inputs and not on the time. If we did not want to make this assumption, we could add T as a parameter to our assumptions.

output O , as well as the corresponding $ok(X, I_1, I_2, O)$. The operations of the gate can then be specified as

fact $andg(X) \wedge ok(X, in1(X), in2(X), out(X))$
 $\Rightarrow out(X, T) = and(in1(X, T), in2(X, T)).$
default $ok(X, I1, I2).$
fact $andg(X) \wedge ab(X, in1(X), in2(X), V)$
 $\wedge V \neq and(in1(X, T), in2(X, T))$
 $\Rightarrow out(X, T) = V.$
conjecture $ab(X, I1, I2, O).$

The first fact says that the output of a normal and gate is the conjunction of the inputs. The second fact says that the output of an abnormal gate (i.e. abnormal for the particular input values) is some value which is different to the conjunction of the inputs.

The main differences between the diagnoses is that Theorist does not need to make assumptions about parts which are not relevant to the diagnosis (we minimise all assumptions, whereas Reiter maximises normality assumptions). By ok we mean that the gate is working normally for the particular inputs being considered. Reiter means (by $\neg ab(X)$) that X is working normally for all inputs. Theorist can have a gate being OK for some inputs and not OK for other inputs.

It is simple to incorporate fault models into Theorist. For example, if we want to say that faulty gates are either stuck at one or stuck at zero (admittedly a very naive assumption), this can be specified by restricting what can be conjectured:

fact $andg(X) \wedge stuck(X, V) \Rightarrow out(X, T) = V.$
conjecture $stuck(X, V).$

Reiter would specify this as

$andg(X) \wedge ab(X) \wedge \neg ab'(X) \Rightarrow stuck1(X) \vee stuck0(X)$
 $stuck1(X) \Rightarrow out(X, T) = 1$
 $stuck0(X) \Rightarrow out(X, T) = 0$

Note that the use of this axiom is very different to the use of the Theorist version. This is only used to say that a gate by default is not broken

because it is not stuck at one or stuck at zero. This is only useful if we can indeed prove that one of these is not the case. For more complicated cases it is easy to imagine a situation where we cannot actually prove that some abnormality does not occur. This is very different to being able to conjecture a fault. Also Reiter's diagnosis does not say that the gate is stuck at one, it just says that the gate is abnormal.

8 Conclusion

In this paper I have tried to present an argument as to why some distinctions are important to make in hypothetical reasoning, and proposed a system which uses these distinctions and have attempted to examine the consequences of these distinctions particularly with respect to prediction and explanation problems.

An important feature of this work is that I have not proposed a new logic in any shape or form. I have tried to be careful in arguing that there are useful ways to use logic and to then consider consequences of these strategies on building AI programs. The success of this can be gauged by finding how many problems go away when logic is viewed in this way as the basis for a hypothetical reasoning system.

Acknowledgements

This work could not have been done without the ideas, criticism, feedback and support of Randy Goebel, Eric Neufeld, Bruce Kirby, Paul Van Arragon, Romas Aleliunas, Scott Goodwin and Denis Gagné. This research was supported under NSERC grant A6260.

References

- [Brown82] J. S. Brown, R. R. Burton and J. de Kleer, "Pedagogical, natural language and knowledge engineering techniques in SOPHIE I, II and III", in D. Sleeman and J. S. Brown (ed.), *Intelligent Tutoring Systems*, Academic Press, New York, pp. 227-282.

- [Chang73] C-L. Chang and R. C-T. Lee, *Symbolic Logic and Mechanical Theorem Proving*, Academic Press, 1973.
- [Charniak85] E. Charniak and D. McDermott, *Introduction to Artificial Intelligence*, Addison-Wesley.
- [Cox87] P. T. Cox and T. Pietrzykowski, *General Diagnosis by Abductive Inference*, Technical report CS8701, School of Computer Science, Technical University of Nova Scotia, April 1987.
- [Davis84] R. Davis, "Diagnostic Reasoning Based on Structure and Behaviour", *Artificial Intelligence* 24, pp. 347-410.
- [Doyle79] J. Doyle, "A Truth Maintenance System", *Artificial Intelligence*, Vol. 12, pp 231-273.
- [de Kleer86] J. de Kleer, "An Assumption-based TMS", *Artificial Intelligence*, Vol. 28, No. 2, pp. 127-162.
- [de Kleer87] J. de Kleer, "Diagnosing Multiple Faults", *Artificial Intelligence*, Vol. 32, No. 1, pp. 97-130.
- [Enderton72] H. B. Enderton, *A Mathematical Introduction to Logic*, Academic Press, Orlando.
- [Gagné87] D. Gagné, *The Multiple Extension Problem Revisited*, in Technical Report CS-87-30, Department of Computer Science, University of Waterloo.
- [Genesereth84] M. R. Genesereth, "The Use of Design Descriptions in Automated Diagnosis", *Artificial Intelligence*, Vol. 24, pp. 411-436.
- [Genesereth87] M. R. Genesereth and N. Nilsson, *Logical Foundations of Artificial Intelligence*, Morgan-Kaufmann.
- [Goebel87] R. G. Goebel and S. D. Goodwin, "Applying theory formation to the planning problem" in F. M. Brown (Ed.), *Proceedings of the 1987 Workshop on The Frame Problem in Artificial Intelligence*, Morgan Kaufmann, pp. 207-232.

- [Goodwin87] S. D. Goodwin, *Representing Frame Axioms as Defaults*, Research Report CS-87-48, Department of Computer Science, University of Waterloo, July 1987, 186 pages.
- [Harman86] G. Harman, *Change in View*, MIT Press.
- [Hayes77] P. J. Hayes, "In Defence of Logic", *Proc. IJCAI-77*, pp. 559-565.
- [Konolige87] K. Konolige, "On the relationship between default theories and autoepistemic logic", *Proc. IJCAI-87*, pp. 394-401.
- [Moore82] R. C. Moore, "The Role of Logic in Knowledge Representation and Commonsense Reasoning", *Proc. AAAI-82*, pp. 428-433.
- [Moore85] R. C. Moore, "Semantical Considerations on Nonmonotonic Logic", *Artificial Intelligence*, Vol. 25, No. 1, pp. 75-94. pp 272-279.
- [Neufeld87a] E. M. Neufeld and D. Poole, "Towards solving the multiple extension problem: combining defaults and probabilities", to appear *Workshop on Reasoning with Uncertainty*, Seattle, July 1987.
- [Neufeld87b] E. M. Neufeld and D. Poole, "Combining Defaults and Probability for Prediction", to be submitted to CSCSI-88.
- [Patil81] R. S. Patil, P. Szolovits and W. B. Schwartz, "Causal understanding of patient illness in medical diagnosis", *Proc. IJCAI-81*, pp. 893-899.
- [Pearl87] J. Pearl, "Embracing Causality in Formal Reasoning", *Proc. AAAI-87*, pp. 369-373.
- [Poole85] D. L. Poole, "On the Comparison of Theories: Preferring the Most Specific Explanation", *Proc. IJCAI-85*, pp.144-147.
- [PGA87] D. L. Poole, R. G. Goebel, and R. Aleliunas, "Theorist: a logical reasoning system for defaults and diagnosis", in N. Cercone and G. McCalla (Eds.) *The Knowledge Frontier: Essays in the Representation of Knowledge*, Springer Verlag, New York, 1987, pp. 331-352; also Research Report CS-86-06, Department of Computer Science, University of Waterloo, 16 pages, February 1986.

- [Poole87a] D. L. Poole, *A Logical Framework for Default Reasoning*, submitted.
- [Poole87b] D. L. Poole (Ed.), *Experiments in the Theorist Paradigm: A Collection of student papers on the Theorist Project*, Research Report CS-87-30, Department of Computer Science, University of Waterloo, May.
- [Popl83] H. E. Popl, Jr., "Heuristic methods for imposing structure on ill structured problems", in P. Szolovits (Ed.), *Artificial Intelligence in Medicine*, AAAS/Westview, Boulder, CO, pp. 119-190.
- [Popper62] K. R. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge*, Basic Books, New York.
- [Quine78] W. V. Quine, and J. S. Ullian, *The Web of Belief*, Random House, New York, Second Edition.
- [Reiter80] R. Reiter, "A Logic for Default Reasoning", *Artificial Intelligence*, Vol. 13, pp 81-132.
- [Reiter87] R. Reiter, "A Theory of Diagnosis from First Principles" *Artificial Intelligence*, Vol. 32, No. 1, pp. 57-96.
- [Weiss78] S. M. Weiss, C. A. Kulikowski, S. Amarel and A. Safir, "A model-based method for computer-aided medical decision making", *Artificial Intelligence* 11, pp. 145-172.

REMITTANCE ADVICE

PRINCETON UNIVERSITY

864657

VOUCHER NUMBER	VOUCHER DATE				GROSS	DISCOUNT	NET
		INVOICE NO.	PURCHASE ORDER				
321606	05-31-88	*			4.00		4.00
					4.00		4.00

*received cheque # 864657
sent reports
JUN 8 1988*

DETACH BEFORE DEPOSIT

65/406

Department of Computer Science
University of Waterloo
Research Reports 1987 (September to December) continued

Report No.	Title	Author	Cost
CS-87-49	Networks for Education at the University of Waterloo	D.D. Cowan S.L. Fenton J.W. Graham T.M. Stepien	2.00
CS-87-50	A Network Operating System For Interconnected LANS With Heterogeneous Data-Link Layers	D.D. Cowan T.M. Stepien R.G. Veitch	2.00
CS-87-51	Project ARIES A Network for Convenient Computing in Education	D.D. Cowan T.M. Stepien	2.00
CS-87-52	Naming of Objects in the Cluster System	F.C.M. Lau J.P. Black E.G. Manning	2.00
CS-87-53	A Study of Distributed Debugging	W.H. Cheung J.P. Black E.G. Manning	2.00
CS-87-54	Defaults and Conjectures: Hypothetical Reasoning for Explanation and Prediction	D.L. Poole	2.00
CS-87-55	Roughly Sorting: A Generalization of Sorting	Y. Igarashi D. Wood	2.00
CS-87-56	A Dynamic Fixed Windowing Problem	R. Klein O. Nurmi T. Ottmann D. Wood	2.00
CS-87-57	Explorations in Restricted Orientation Geometry	G.J.E. Rawlins	5.00
CS-87-58	A New Measure of Presortedness	V. Estivill-Castro D. Wood	2.00
CS-87-59	A Logical Framework for Default Reasoning	D.L. Poole	2.00
CS-87-60	On Consistent Equiarea Triangulations	R.B. Simpson	2.00

CS-87-61	Amalgamating Functional and Relational Programming through the Use of Equality Axioms	K. Yukawa	5.00
CS-87-62	Comparison of the Single Phase and Two Phase Numerical Model Formulation For Saturated-Unsaturated Ground Water Flow (in preparation)	P. Forsyth	2.00
CS-87-63	On the Existence of Speed-Independent Circuits	C-J. Seger	2.00
CS-87-64	Logic-based Program Transformation (in preparation)	M.H.M. Cheng M.H. van Emden P.A. Strooper	?
CS-87-65	An Algebra for Nested Relations	V. Deshpande P.-A. Larson	2.00
CS-87-66	On The Average Case of String Matching Algorithms	R. Baeza-Yates	2.00
CS-87-67	Perceptual Reasoning: A Logical Foundation for Computer Vision	J.D.Denis Gagne	2.00
CS-87-68	Searching with Uncertainty	R. Baeza-Yates J.C. Culberson G.J.E. Rawlins	2.00

If you would like to order any reports please forward your order, along with a cheque or money order payable to the **Department of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1** to the **Research Report Secretary**.

Please indicate your current mailing address.

Att: Laura Hawkins
Princeton University
Cognitive Science Lab
221 Nassau St. 1st floor
Princeton, NJ 08542

Remitter:



DANSK DATAMATIK CENTER

Lundtoftevej 1 c

DK-2800 Lyngby

Date:

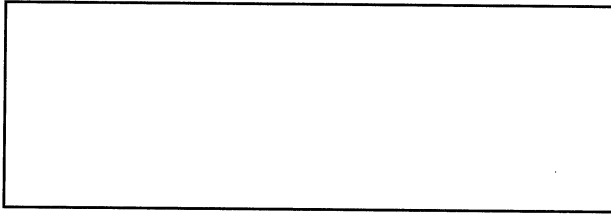
No.: **837770**

Enclosed please find cheque issued by

SPAREKASSEN **sds**

for

in cover of



- ☐ Please return enclosed invoice duly receipted
- ☐ Please acknowledge receipt
- ☐ No acknowledgment required

Yours faithfully,



University of Waterloo
Dept. of Computer Science
Waterloo,
Ontario, N2L 3G1
USA

Lyngby, 1988-05-31

Dear Madam,

Please send me a copy of the two reports marked with a X on the enclosed papers.

Enclosed please find a cheque for USD 4,0.

Thank you.

Yours faithfully

Dansk Datamatik Center

Kirsten H. Karlsen
Kirsten H. Karlsen

*received
cheque & sent
reports
June 17/88*

Please mail to:

Dansk Datamatik Center
Att. Kirsten H. Karlsen
Lundtoftevej 1.C
2800 Lyngby
DENMARK

IND G A L I
03 FEB. 1988

Department of Computer Science
University of Waterloo
Research Reports 1987 (September to December) continued

Report No.	Title	Author	Cost
CS-87-49	Networks for Education at the University of Waterloo	D.D. Cowan S.L. Fenton J.W. Graham T.M. Stepien	2.00
CS-87-50	A Network Operating System For Interconnected LANS With Heterogeneous Data-Link Layers	D.D. Cowan T.M. Stepien R.G. Veitch	2.00
CS-87-51	Project ARIES A Network for Convenient Computing in Education	D.D. Cowan T.M. Stepien	2.00
CS-87-52	Naming of Objects in the Cluster System	F.C.M. Lau J.P. Black E.G. Manning	2.00
CS-87-53	A Study of Distributed Debugging	W.H. Cheung J.P. Black E.G. Manning	2.00
X CS-87-54	Defaults and Conjectures: Hypothetical Reasoning for Explanation and Prediction	D.L. Poole	2.00
CS-87-55	Roughly Sorting: A Generalization of Sorting	Y. Igarashi D. Wood	2.00
CS-87-56	A Dynamic Fixed Windowing Problem	R. Klein O. Nurmi T. Ottmann D. Wood	2.00
CS-87-57	Explorations in Restricted Orientation Geometry	G.J.E. Rawlins	5.00
CS-87-58	A New Measure of Presortedness	V. Estivill-Castro D. Wood	2.00
X CS-87-59	A Logical Framework for Default Reasoning	D.L. Poole	2.00
CS-87-60	On Consistent Equiarea Triangulations	R.B. Simpson	2.00

Printing Requisition / Graphic Services

1256

1. Please complete unshaded areas on form as applicable.
2. Distribute copies as follows: White and Yellow to Graphic Services. Retain Pink Copies for your records.
3. On completion of order the Yellow copy will be returned with the printed material.
4. Please direct enquiries, quoting requisition number and account number, extension 3451.

TITLE OR DESCRIPTION

C'S-87-54

DATE REQUISITIONED

Dec. 14/87

DATE REQUIRED

11/01/88

ACCOUNT NO.

11261626041

REQUISITIONER - PRINT

S. DEANGELIS

PHONE

2172

SIGNING AUTHORITY

Sue DeAngelis / D. Poole

MAILING INFO -

NAME

S. DEANGELIS

DEPT.

C.S.

BLDG. & ROOM NO.

MVC 6081E

☒ DELIVER
☐ PICK-UP

Copyright: I hereby agree to assume all responsibility and liability for any infringement of copyrights and/or patent rights which may arise from the processing of, and reproduction of, any of the materials herein requested. I further agree to indemnify and hold blameless the University of Waterloo from any liability which may arise from said processing or reproducing. I also acknowledge that materials processed as a result of this requisition are for educational use only.

NUMBER OF PAGES 50 NUMBER OF COPIES 50

TYPE OF PAPER STOCK

☒ BOND ☐ NCR ☐ PT. ☒ COVER ☐ BRISTOL ☒ SUPPLIED ☐

PAPER SIZE

☒ 8 1/2 x 11 ☐ 8 1/2 x 14 ☐ 11 x 17 ☐

PAPER COLOUR

☒ WHITE ☐ ☒ BLACK ☐

PRINTING

☐ 1 SIDE ☐ PGS. ☒ 2 SIDES ☐ PGS. FROM ☐ TO ☐

BINDING/FINISHING

☒ COLLATING ☒ STAPLING ☐ HOLE PUNCHED ☐ PLASTIC RING

FOLDING/PADDING

CUTTING SIZE

Special Instructions

With fronts & backs enclosed.

COPY CENTRE

OPER. NO. ☐ BLDG. ☐ MACH. NO. ☐

DESIGN & PASTE-UP

OPER. NO. ☐ TIME ☐ LABOUR CODE ☐ D01
☐ D01
☐ D01

TYPESETTING

QUANTITY

PAP000000 T01
PAP000000 T01
PAP000000 T01

PROOF

PRF
PRF
PRF

NEGATIVES

QUANTITY

OPER. NO.

TIME

LABOUR CODE

F L M ☐ ☐ ☐ ☐ ☐ ☐ C01
F L M ☐ ☐ ☐ ☐ ☐ ☐ C01
F L M ☐ ☐ ☐ ☐ ☐ ☐ C01
F L M ☐ ☐ ☐ ☐ ☐ ☐ C01
F L M ☐ ☐ ☐ ☐ ☐ ☐ C01

PMT

P M T ☐ ☐ ☐ ☐ ☐ ☐ C01
P M T ☐ ☐ ☐ ☐ ☐ ☐ C01
P M T ☐ ☐ ☐ ☐ ☐ ☐ C01

PLATES

P L T ☐ ☐ ☐ ☐ ☐ ☐ P01
P L T ☐ ☐ ☐ ☐ ☐ ☐ P01
P L T ☐ ☐ ☐ ☐ ☐ ☐ P01

STOCK

☐ ☐ ☐ ☐ ☐ ☐ 001
☐ ☐ ☐ ☐ ☐ ☐ 001
☐ ☐ ☐ ☐ ☐ ☐ 001
☐ ☐ ☐ ☐ ☐ ☐ 001

BINDERY

R N G ☐ ☐ ☐ ☐ ☐ ☐ B01
R N G ☐ ☐ ☐ ☐ ☐ ☐ B01
R N G ☐ ☐ ☐ ☐ ☐ ☐ B01
M I S 000000 B01

OUTSIDE SERVICES

\$ ☐ COST
TAXES - PROVINCIAL ☐ FEDERAL ☐ GRAPHIC SERV. OCT. 85 482-2

86842

1. Please complete unshaded areas on form as applicable.
2. Distribute copies as follows: White and Yellow to Graphic Services. Retain Pink Copies for your records.
3. On completion of order the Yellow copy will be returned with the printed material.
4. Please direct enquiries, quoting requisition number and account number, to extension 3451.

TITLE OR DESCRIPTION CC-27-54 Give Kim 55 copies			
DATE REQUISITIONED Oct. 21/89		DATE REQUIRED ASAP	
ACCOUNT NO. 11261626041		SIGNING AUTHORITY Kim Shuerich	
REQUISITIONER-PRINT K. Shuerich		PHONE 2192	
MAILING NAME INFO - SUP materials		DEPT. CC	
BLDG. & ROOM NO. MC 6021E		<input checked="" type="checkbox"/> DELIVER <input type="checkbox"/> PICK-UP	

Copyright: I hereby agree to assume all responsibility and liability for any infringement of copyrights and/or patent rights which may arise from the processing of, and reproduction of, any of the materials herein requested. I further agree to indemnify and hold blameless the University of Waterloo from any liability which may arise from said processing or reproducing. I also acknowledge that materials processed as a result of this requisition are for educational use only.

[illegible]