Artificial Intelligence Review, (1989), **3**, 129–158

Abductive reasoning in multiple fault diagnosis

T. Finin & G. Morris*,
*Paoli Research Center, Unisys, Paoli, PA, USA and *AI Laboratory, Internal Revenue Service, Washington, DC, USA*

Abstract. Abductive reasoning involves generating an explanation for a given set of observations about the world. Abduction provides a good reasoning framework for many AI problems, including diagnosis, plan recognition and learning. This paper focuses on the use of abductive reasoning in diagnostic systems in which there may be more than one underlying cause for the observed symptoms. In exploring this topic, we will review and compare several different approaches, including Binary Choice Bayesian, Sequential Bayesian, Causal Model Based Abduction, Parsimonious Set Covering, and the use of First Order Logic. Throughout the paper we will use as an example a simple diagnostic problem involving automotive troubleshooting.

1 Introduction

The technologies of knowledge-based expert systems have been applied to many different types of problems. Diagnosis has been one of the earliest applications areas as well as being one of the most important and interesting. One attempt to formalize diagnosis is as *Abduction* — reasoning from a set of observations about the world to a hypothesis that explains or accounts for the observations. This paper focuses on the use of abductive reasoning in diagnostic systems in which there may be more than one underlying cause for the observed symptoms. In exploring this topic, we review and compare several different approaches, including Binary Choice Bayesian, Sequential Bayesian, Causal Model Based Abduction, Parsimonious Set Covering, and the use of First Order Logic. Throughout the paper we use, as an example, a simple diagnostic problem involving automotive troubleshooting adapted from (Weiss & Kulikowski, 1984).

Numerous expert systems have been developed for diagnostic reasoning, many of them in the medical area. Some of the earliest successful systems were rule-based deductive programs, like *MYCIN* (Shortliffe, 1976, Buchanan & Shortliffe, 1984). A common criticism of these pioneering efforts was their handling of situations where more than one disease was needed to explain correctly all the observed symptoms. Each of the systems we discuss in this paper held as a major design objective the correct handling of multiple faults in diagnostic problems. Collectively, they represent most of the current approaches.

In this section we introduce the concept of abductive reasoning and present a simple problem to be used as a running example in the remainder of the paper. Section 2 describes the five systems objectively. Section 3 addresses the comparative merits of these approaches, including any theoretical weaknesses of the real-world implementations. Section 4 summarizes the major issues along which the approaches differ and describes an emerging consensus for the formalization of the diagnostic process.

Definition of abductive reasoning

Although many diagnostic systems are not strongly tied to first-order logic, the diagnostic process is clearly an example of *abductive reasoning*. Pople (1973) defines abductive logic within the realm of first order logic with the following schema:

I	Major Premise (rule)	$\forall x[P(x) \Rightarrow Q(x)]$
II	Minor Premise (case)	$P(a)$
III	Conclusion (fact)	$Q(a)$

Deductive logic involves reasoning from a rule (I) and a case (II) to a conclusion (III). If we know the rule and also that P(a) is true, we may conclude Q(a). A deducive conclusion is certain if its bases (I and II) are sound. *Inductive* logic involves reasoning from a case and a conclusion toward a rule. If we see that P(a) is true and also observe that Q(a) is true, we may hypothesize that 'perhaps all things P are also Q'. *Abductive* logic is the third possibility — it involves reasoning from a fact (III) and a rule (I) toward a case (II). If we observe that Q(a) is true, and we know the rule 'all things P are Q', we may hypothesize that 'perhaps *a* is P'. Neither inductive nor abductive reasoning leads to certainty; we must hypothesize, and there may be several competing hypotheses that could be logically correct. This is the nature of most diagnostic tasks.

Note that abduction is different from backward chaining, although both could be called 'using a rule backwards'. In backwards chaining, the goal 'prove that *a* is Q' gives rise to a sub-goal of 'prove that *a* is P'. If the sub-goal can be achieved, then one may *deduce* Q(a). In abductive reasoning, on the other hand, one formulates hypotheses to explain symptoms which are not goals but observable facts. We already know that '*a* is a Q' and the task is to abduce why so that *a* can be cured of disorder Q.

Another, more significant, difference between abduction and deductive 'backward chaining' has to do with the *causal* nature of abductive rules. Although one could define abduction syntactically, as we have done above, this does not really capture the sense the word today, as it is used in the AI community. Abduction requires that the 'rules' capture *causal* relationships in order for the conclusions to be true explanations. For example, one might find the following implication in expert system to troubleshoot automotive engine:

$$Knock(X) \Rightarrow Bad\ Timing(X)$$

This rule might capture the diagnostic rule that engine knock is a symptom which implies that the engine's timing is bad. However, this rule cannot be used to generate the fact *Knock(car 23)* as an explanation for the observation *Bad Timing(car 23)*. A more troubling example could be obtained from the logically correct rule

$$and(X,Y) \Rightarrow X,$$

which would lead to explaining any observation O with the explanation *and* $(2+2=4,\ O)$.

Logical implication and the causal relation are not identical. The fact that they are closely related and that implications are a natural way to express causality adds to the confusion. We must keep in mind that, in addition to our syntactic definition of abductive reasoning, we will require that the 'rules' over which it reasons must express causal relationships.

Other applications of abductive reasoning

Abductive reasoning is a useful approach to many other AI problems as well. Whenever we are presented with a set of observations about the world and are charged with devising a hypothesis which will explain them, we are dealing with an abductive problem. This general scenario matches a number of standard problems, a few of which we will briefly mention.

There has been a great deal of research in the last 10 years aimed at providing co-operative interfaces to systems such as expert systems (Pollack *et. al.*, 1982; Finin *et. al.*, 1986; Pollack, 1986), database retrieval systems (Kaplan, 1982; Carberry, 1987), and in a more general question-answering context (Allen, 1982). Truly co-operative systems need to be able to address their user's underlying goals in using the system. In order to do this, it is necessary to recognize the user's previous queries and statements as forming a plan to achieve some appropriate domain goal.

A similar problem arises in the context of providing intelligent help and advice. In order to provide the information a user needs, it is necessary to have, among other things, a model of what he is trying to accomplish. Again, this involves fitting a user's recent actions into a coherent plan to accomplish some relevant domain goal. Examples of such intelligent help systems include *The Macsyma Advisor* (Genesereth, 1979), which helped a Macsyma user recover from an error state, and *Wizard* (Shrager & Finin, 1982; Finin, 1983), a system which volunteered advice on better ways to use the *VAX/VMS* operating system.

Understanding some extended discourse involving the actions of people, such as a newspaper article or a story, is another problem which requires one to reason abductively from a set of actions being performed to a hypothesis which would explain them. Schank and his colleagues at Yale University have made an extensive study of this problem (Schank & Abelson, 1977). Wilenski has studied the interplay between planning, plan recognition and behaviour understanding in

(Wilenski, 1983). Kautz has proposed a new, richer approach to general plan recognition which is directly based on abduction (Kautz, 1985).

In summary, the world presents us with abductive problems of many kinds. Sometimes we must understand the internal state of an artifact from a small set of symptoms it presents. We are constantly trying to understand the internal goal structure of other people based on their observed behaviour. The state of some subset of the physical world (e.g. the Amazon Rain Forest Ecology or the U.S. Economy) must be inferred from a set of observable indicators. All of these problems can be formalized as Abductive Reasoning problems.

A running example

To help to understand each of the diagnostic systems described, a simple example diagnostic problem will be used throughout this paper. It is taken from the domain

Table 1. Imaginary knowledge base for auto repair

Symptoms	*Intermediate States*	*Disorders*
Overheating	Late-firing knock, incomplete combustion	Bad timing
	Excess cooling system pressure, late circulation on warm-up	Bad thermostat
	Excess cooling system pressure	Clogged radiator
Poor mileage	Excess trans/clutch wear	Trans/clutch slipping
	Excess brake pad wear, delayed braking loss	Dragging brake
	Late-firing, knock incomplete combustion	Bad timing
	Too rich/lean fuel mix	Bad carburettor chip
	Too rich fuel (warm)	Bad auto choke
	Gas odour, fire hazard	Gas line leak
Poor power	Late-firing, knock, incomplete combustion	Bad timing
	Too rich/lean fuel mix	Bad carburettor chip
	Excess trans/clutch wear	Trans/clutch slipping
	Excess brake pad wear, Delayed braking loss	Dragging brake
Stalls when cold	'Cold idle' too slow too lean fuel mix	Bad carburettor chip
	Too lean fuel mix	Bad auto-choke
	'Cold idle' too slow	Bad temp sender
Stalls when hot	Unstable idle too rich fuel at idle	Bad carburettor chip
Dead battery	Slow loss of charge	Bad alternator
	Slow loss of charge	Bad volt regulator
	Unable to recharge	Major short
		Bad battery
No headlights	Other lights OK	Bad fuse
	Rapid loss of charge	Short in lights

of automobile troubleshooting, loosely after Weiss & Kulikowski (1984). Table 1 presents some imaginary facts about engine performance.

In this table, *Symptoms* are behaviours that can be easily observed, and would probably be noticed by the car owner. *Disorders* are the initial causes of the engine malfunctions. These are also the conditions that must be repaired. *Intermediate states* are conditions inside the engine which are caused by the disorders, and in turn cause other intermediate states or symptoms. Typically, the intermediate states are difficult to observe directly.

The general diagnostic problem in this domain could then be described as follows. We are given a set of initial or *presenting* symptoms, such as (*bad-overheating*) and (*poor-mileage*) and desire to determine one or more hypotheses which could account for all of the observed symptoms. Each hypothesis is a set of disorders, such as (*bad-thermostat*) or (*major-short, bad-battery*). Besides accounting for all of the known symptoms, a hypothesis may also predict other symptoms which have not yet been observed. These predictions provide a way to test the validity of a hypothesis by making the observation and seeing if the result is consistent with the prediction.

2 Five different approaches

'This section presents, in some detail, five important approaches to diagnosis which employ abductive reasoning. Three of them represent fully implemented diagnostic tools which contain a mixture of pragmatic domain-inspired heuristics and domain-independent theoretical methods. Two are long-term studies, with several generations of refinement to their designs and their underlying formal models. To present each system *in toto* would be interesting, but would involve many details and issues which distract from the focus on handling of multiple-fault diagnosis. Therefore, the following descriptions are restricted to the basic model or underlying theory and the specific methods used to address multiple faults.

Binary-choice Bayesian abduction

Ben-Bassat *et. al.* have developed the *MEDAS* (Medical Emergency Decision Assistance System) program which is based rather strictly on Bayesian statistics (Ben-Bassat *et. al.* 1980). It is a medium-sized system, covering 50 high-level disorders using nearly 600 symptoms. *MEDAS* is used for initial diagnosis and assessment of life-threatening potential in a hospital emergency room.

The medical knowledge base is elicited and stored in a frame-like format organized by disorder (see Table 2a). The critical knowledge consists of the following estimates for each disorder D_i and each symptom S_j:

$\mathbf{P(D_i)}$ — the prior probability of disorder D_i,
$\mathbf{P(S_j \mid D_i)}$ — the conditional probabality of symptom S_j when disorder D_i is present,
$\mathbf{P(S_j \mid \bar{D}_i)}$ — the conditional probablity of symptom S_j when disorder D_i is absent.

These parameters are estimated using an eight-level interval scale shown in Table 2. Pathognomonic symptoms (i.e. symptoms which are sure indicators of a specific

Table 2. A MEDAS-style frame for the auto repair knowledge base

	LEGEND	
Symbol	*Meaning*	*Probability*
M	Must	1.0
VP	Very probable	0.90–1.0
P	Probable	0.75–0.90
F	Frequent	0.50–0.75
S	Sometimes	0.25–0.50
R	Rare	0.10–0.25
VR	Very rare	0–0.10
N	Must not	0

Table 2.a. Conditional Probabilities

	DISORDER PATTERN Bad Automatic Choke			
Severity Level (1–5) = 3 Prior Probability = Rare			*Conditional Probability* *Disorder is*	
Item	*Feature Name*	*Cost*	*Present*	*Absent*
1	Poor mileage	1	F	S
2	Poor power	1	S	S
3	Stalls cold	2	P	R
4	H/O Hard starting	2	P	R
5	S/O Bad carburettor chip	2	F	S
.				
.				
.				
12	Stalls hot	3	F	R
.				
.				
.				

disorder) are reflected as certainties. When a symptom is seen **iff** the disorder is present, then $P(S_j \mid D_i) = 1$ and $P(S_j \mid \bar{D}_i) = 0$; when the symptom is a sure indicator, but may not always be present, then only $P(S_j \mid \bar{D}_i) = 0$.

Note that in *MEDAS*, a symptom is essentially a proposition that is either true (if the symptom is present) or false (if the symptom is absent). Symptoms which refer to a value on a scale (e.g. *the patient's pulse rate*) must be converted to range memberships (e.g. *pulse* ≤ 110). Note that in Table 2.a some of the symptoms are of the type 'S/O' or 'Setting of' some other disorder. These 'symptoms' actually reflect the interations of disorders; in this case a Bad Choke is frequently associated with coincident Bad Carburettor Chips. This mechanism allows intermediate pathological states (such as 'fuel mix too lean' in this example) to be recognized and reasoned about. 'H/O' or 'History of' symptoms refer to disorders noted in the car's (or patient's) history. For its emergency room setting, the *MEDAS* knowledge base also explicitly records the life-threatening potential of each disorder.

MEDAS is a Binary-Choice system; that is, it calculates for each individual disorder the posterior probability of its presence or absence given the collection of symptoms which have been reported so far. No attempt is made to deal with combinations of disorders, *per se*. For each disorder D_i, the conditional probability of its presence given the observation of n specific symptoms $s_1...s_n$ is given in Table 3.

By restricting each calculation to the binary choice between the presence or absence of a disorder, Ben-Bassat makes good on the Bayesian assumption that the list of possible disorders is mutually exclusive and exhaustive. The companion assumption — that all symptoms are independent — does not hold, but he deals with *gross* violations of the assumption in the following manner. A group of highly interdependent symptoms (e.g. 'stalls when hot' and 'stalls when cold') for any one disorder are marked, and observation of the first symptom in the group results in augmentation of the disorder's probability. When other symptoms from the interdependent group are observed, they are noted as present but the probability of the disorder is not revised.

One purely heuristic element of *MEDAS* is the identification of 'primary' symptoms, those which are considered the hallmarks of a particular disorder. These are not pathognomonic symptoms, and hence have no special statistical significance. But they are so strongly associated with the disorder by physicians (and so easily confirmed) that physicians will not accept *MEDAS*' diagnosis unless the presence of all primary symptoms has been confirmed — even if the probability of the disorder is already very high. Thus, *MEDAS* will not announce its diagnosis until all primary symptoms for an indicated disorder have been investigated.

The diagnostic routine for *MEDAS* is very straightforward. The initial group of reported symptoms is used to identify candidate disorders. The disorder with the highest probability (or a less likely disorder with large life-threatening potential) is chosen to guide a question–generation phase. A symptom which has not yet been reported is considered for investigation based upon its cost (in dollars, delay and discomfort to the patient) and its potential contribution to eliminating or confirming this disorder. After each new batch of symptoms is reported, the probabilities of *all* disorders (not just candidates) are updated, and the cycle is repeated.

The result of any one cycle is a ranked list of disorders and their (binary) probabilities of being present. No effort is made to determine how many disorders are present. All decisions are made by the user/physician, including the decision to stop the diagnostic cycle. There is also provision for focusing the diagnosis on one area or disorder — not necessarily highly ranked by *MEDAS* — identified by the user as the most important.

Consider our automobile troubleshooting example. Assume that only three symptoms are reported initially; *poor mileage, no headlights* and *stalls cold. Poor Mileage* (pm) is associated with six disorders, per Fig. 1, *stalls cold* (sc) with three, and *no headlights* (nh) with two. The posterior probability of each one of them, given the presence of these symptoms, is calculated per the equation in Table 3. For the disorder *bad choke* (bc)[1], this calculation yields a conditional probability 0.625 after round 1 as is shown in Table 4.

Note that because headlight symptoms have no association with choke problems, that symptom has no impact on the probability of a bad choke. After all disorders have been updated, assume that *bad carburettor chip* is the best looking hypothesis, with probability 0.75, and *bad choke* is second with no other disorder seeming likely. Assume that, after looking at the ordered list of hypotheses, the user decided to instruct *MEDAS* to pursue the choke alternative rather than the carburettor. So *MEDAS* evaluates the potential contribution of each other symptom in Table 2 against the cost of acquiring information about it. *Poor power* has a cost of 1, but its presence is not a strong indicator of bad choke, so *History of Hard Starting* is chosen despite its cost of 2. The question is posed to the user, and the answer restarts the cycle.

Table 3. The conditional probability of a disease given a set of symptoms

$$P(D_i \mid s_1 \ldots s_n) = \frac{P(D_i)f_i(s_1) \ldots f_i(s_n)}{P(D_i)f_i(s_1) \ldots f_i(s_n) + (1 - P(D_i))\bar{f}_i(s_1) \ldots \bar{f}_i(s_n)}$$

$$f_i(s_j) = \begin{cases} P(s_j \mid D_i) & \text{if } s_j \text{ is present} \\ 1 - P(s_j \mid D_i) & \text{if } s_j \text{ is absent} \end{cases}$$

$$f_i(s_j) = \begin{cases} P(s_j \mid \bar{D}_i) & \text{if } s_j \text{ is present} \\ 1 - P(s_j \mid \bar{D}_i) & \text{if } s_j \text{ is absent} \end{cases}$$

Table 4. Probability of *bad choke* after round one

$$\begin{aligned} P(bc)_{Round\ 1} &= \frac{P(bc) \cdot P(pm \mid bc) \cdot P(sc \mid bc)}{P(bc) \cdot P(pm \mid bc) \cdot P(sc \cdot bc) + (1-P(bc)) \cdot P(pm \mid \overline{bc}) \cdot P(sc \mid \overline{bc})} \\ &= \frac{(0.1)(0.5)(0.75)}{(0.1)(0.5)(0.75) + (0.9)(0.25)(0.1)} \\ &= 0.625 \end{aligned}$$

INTERNIST — a sequential Bayesian approach

A landmark effort to diagnose multiple simultaneous disorders is the *INTERNIST* system created primarily by Pople & Myers (Pople 1977; Pople 1982). It, too, is based on the Bayesian formula for determining the posterior probability of a disorder given a group of symptoms. Like *MEDAS*, *INTERNIST* acknowledges that the diseases considered by it are not a mutually exclusive and exhaustive list. However, *INTERNIST* deals with this by using an alternative formulation of Bayes' Theorem (Charniak, 1983), which only assumes the independence of symptoms in general and the independence of symptoms given the existence of some disease:

$$P(D_i \mid s_1 \ldots s_n) = \frac{P(D_i) \cdot P(s_1) \ldots P(s_n \mid D_i)}{P(s_1) \ldots P(s_n)} \quad (1).$$

INTERNIST records its medical knowledge base in the form of relations on diseases and symptoms. The *Evokes* relation records the evoking strength of the symptom for each disease; it is analogous to $P(D \mid s)$, but measured on a scale of 0–5. Similarly, the *Manifests* relation records $P(s \mid D)$ on a 1–5 scale. For example, the automobile data in Table 1 would be recorded in *INTERNIST* as in Table 5.[2]

Like *MEDAS*, *INTERNIST* updates the probability of each disorder based on the symptoms which are initially observed. Based upon the initial data, the highest-ranked (single-disorder) hypothesis is used to form a 'decision set'. Each disorder which might account for the same symptoms as the highest-ranked disorder (or any subset of those symptoms), and whose probability exceeds a threshold, is included in the set. *INTERNIST* then focuses upon differential diagnosis within that decision set, i.e. it attempts to rule out all disorders but one. This is done by requesting symptom observations or laboratory test results from the user/physician. New information is requested based on its cost and its value in distinguishing among the disorders in the decision set.

Table 5. *INTERNIST*-style representation of imaginary auto repair data

EVOKES relation for poor mileage		*MANIFESTS* relation for bad automatic choke	
Disorder	*Strength*	*Symptom*	*Strength*
Trans/clutch slipping	2	Poor mileage	3
Dragging brake	1	Poor power	2
Bad timing	3	Stalls cold	4
Bad carburettor chip	3	H/O hard starting	4
Bad automatic choke	2	Stalls hot	3
⋮	⋮	⋮	⋮

When the new information is obtained, it is used to update the probabilities of *all* disorders, not just members of the decision set. The highest-ranking disorder is used to form a (perhaps new) decision set, and the cycle is repeated. Depending upon the size of the decision set, *INTERNIST* will choose one of the following strategies:

(a) The *Ruleout* strategy is used to pare down a list of more than five disorders. It pursues information which could rule out one or more candidate disorders.

(b) The *Discriminate* strategy is used when two to four candidates remain. Information is sought which can best discriminate between the two top candidates.

(c) The *Narrow* strategy is identical to Discriminate, when invoked on more than four contenders. This is done when no helpful questions can be found to rule out a contender without resorting to intrusive tests.

(d) The *Pursuing* strategy is used when a decision set has been resolved to a single disorder. It calls for confirmatory information until the separation of the leader from the next most likely contender exceeds double the threshold value.

When this sub-diagnosis is completed, the winning candidate is recorded and all symptoms encompassed by the decision set are marked 'explained'. The symptoms, and all disorders in the decision set, are removed from further consideration. However, to reflect the interactions of disorders, any hypotheses for the remaining symptoms, which involve a disorder that is associated with the winning candidate, will receive additional points in the hypothesis ranking. The cycle is repeated until all symptoms have been explained or there are no more candidate disorders. *INTERNIST*'s final product is a 'most probable' set of disorders which explain all the observed symptoms.

In the automobile example, again assume that three symptoms are reported initially: *poor mileage, no headlights*, and *stalls when cold*. According to Table 5, *poor mileage* evokes both *bad carburettor chip* and *bad timing* with an evoking strength of 3. It evokes *transmission/clutch slipping* and *bad choke* with strength 2. Assume that *stalls when cold* evokes *bad choke* with strength 2, *bad carburettor chip* and *bad temperature sender* with strength 1, and that *no headlights* evokes *bad fuse* with strength 3 and *short in lights* with strength 1. Evoking strengths in *INTERNIST* are on a base-2 log scale (e.g. strength 4 is twice as strong as 3) so the combined effect of the symptoms is shown in Table 6.

Table 6. *INTERNIST*-style representation of the sample problem

	Evoking strength of			
Disorder	*pm*	*sc*	*nh*	*combined*
Bad choke	2	2	–	3
Bad timing	3	–	–	3
Bad carburettor chip	3	1	–	3^+
Trans/clutch	2	–	–	2
Bad sender	–	1	–	1
Dragging brake	1	–	–	1
Bad fuse	–	–	3	3
Short in lights	–	–	1	1

The highest-ranked hypothesis after initial input is *bad carburettor chip* and the top six disorders can explain the same symptoms or a proper subset of them, so the initial decision set consists of those six disorders. Following the *Ruleout* strategy, *INTERNIST* seeks a question which can potentially eliminate one of the candidates. It asks the user whether there is any evidence of *poor power*, looking to eliminate (or stregnthen) the low-ranked *dragging brake* and *transmission/clutch slippage* candidates. The user responds that power is normal. In light of this new information, those two candidates drop from contention, allowing *INTERNIST* to switch to a *Discriminate* strategy which concentrates on the two top-ranked candidates. It now asks about a history of hard starting, and the affirmative answer makes *bad choke* the new top-ranked candidate. After confirming that (partial) diagnosis, *INTERNIST* will record *bad choke* as the winner, and proceed to the remaining symptoms. If any possible cause of *no headlights* were related to *bad choke*, it would receive bonus points in the scoring of subsequent rounds. The final product of *INTERNIST would be a two-disorder diagnosis covering all the reported symptoms.*

ABEL — a non-Bayesian, causal model approach

Patil has built a system for acid-base and electrolyte disorder diagnosis, *ABEL*, which does not use Bayesian statistics at all (Patil, 1981). For our purposes, a key element of Patil's design is that disorder hypotheses are not considered sequentially. Rather, *ABEL* generates hypotheses which are sets of disorders so that each hypothesis can explain all of the observed symptoms. The general preference for parsimonious hypotheses is employed to favour a smaller hypothesis over a larger one. The mechanism for linking symptoms and disorders is not conditional probabilities but a *causal model* which reflects the database of medical knowledge.

Unlike *INTERNIST* or *MEDAS*, *ABEL* uses scalar values of symptoms (like 'air/fuel ratio = 500') and its causal model predicts the magnitudes of symptoms (or manifestations) caused by any one disorder. This allows easy combination of the effects of different disorders. For example, the disorder *bad temperature sender* can cause *air/fuel ratio* values to remain normal (say 600) when the engine is cold. The disorder *bad carburettor chip* could cause the ratio to be 400 ordinarily. Thus, the combination of the two disorders is required to explain fully an observed ratio of 400 with cold engine. However, a third disorder with opposite effects could partially or wholly mask this symptom.

The basic data structure of *ABEL* is a Patient-Specific Model (PSM), composed of relevant fragments of the complete *causal model*. This model (Fig. 1) predicts, for instance, that *low air/fuel ratio* may cause *poor mileage*, but that in turn must be caused by some other factor. Further, the model states that an air/fuel ratio up to 20% too high may be explained by *slow choke release*, but if the ratio is higher than that some other independent cause for it must be presumed in the PSM. To continue the automobile example, assume again the symptoms *poor mileage*, *no headlights*, and *stalls when cold* have been reported. These are instantiated in a PSM (by giving them instance numbers — see the circled nodes in Fig. 2) and matched with the causal model. Any instance which can be explained in terms of other reported (or soundly inferred) findings is marked 'accounted'; otherwise it is 'unaccounted'.

Note that *ABEL*'s domain — acid-base and electrolyte imbalances — facilitates a detailed, quantified model of the operative principles; it is a relatively well-understood area of medicine. The basic diagnostic cycle consists of the following steps:

1 *Presenting complaints:* The initial symptoms are analysed (serum analysis and the initial complaints in this domain). A small set of initial PSMs are created and added to the list of causal hypotheses (the CH-list).
2 *Rank ordering hypotheses:* All PSMs in the CH-list are scored for the quality of explanation they provide for the patient's illness. The leading one or two of these PSMs are selected as possible explanations.

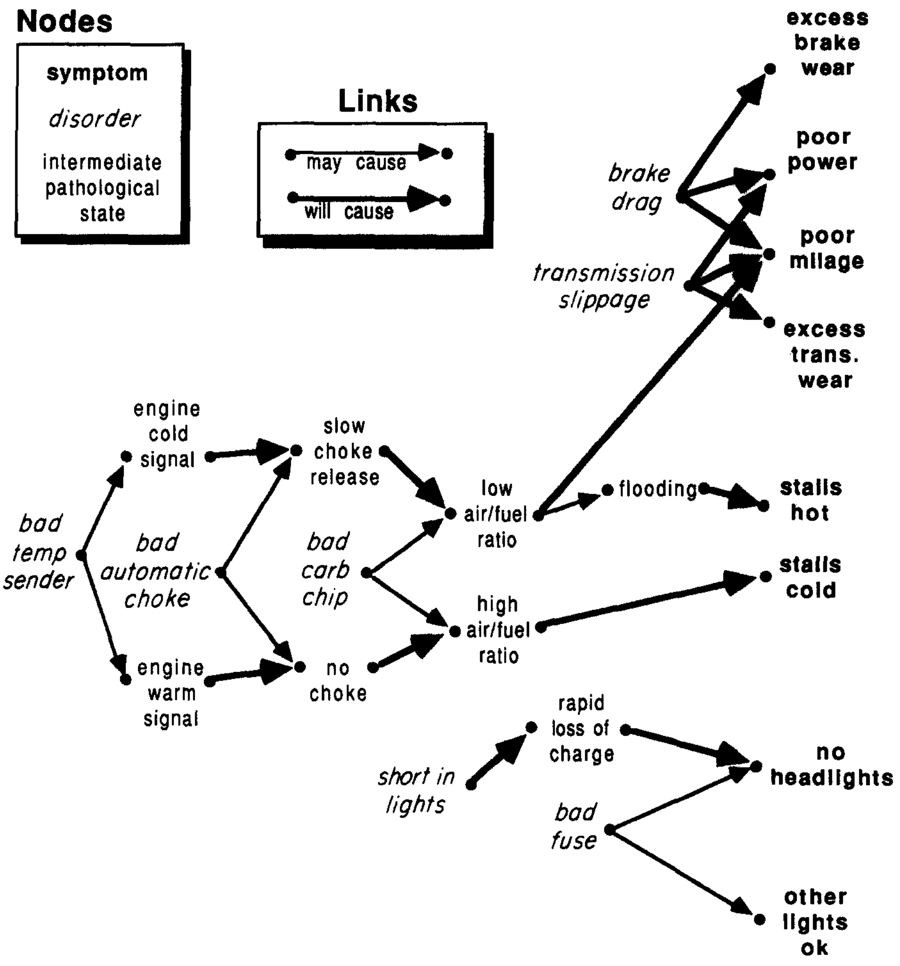

Fig. 1 An *ABEL*-style causal model for the auto repair knowledge base.

3 *Computing diagnostic closure:* Diagnostic Closures [DCs] for the selected PSMs are computed and disease hypotheses in each DC are scored.

4 *Termination:* If the diagnostic closures for all PSMs are null, or if some PSM provides a complete and coherent account for the patient's illness, then the current phase of the diagnosis is complete.

5 *Diagnostic information gathering:* Based on the number of DCs (i.e. the PSMs selected in Step 2), a top level confirm or differentiate goal is formulated. Using diagnostic strategies, this goal is successively decomposed into simpler subproblems until individual questions are formulated.

6 *Re-structuring the PSM:* If Step 5 results in any new finding being known, then that finding is incorporated into each of the PSMs by extending the structure of

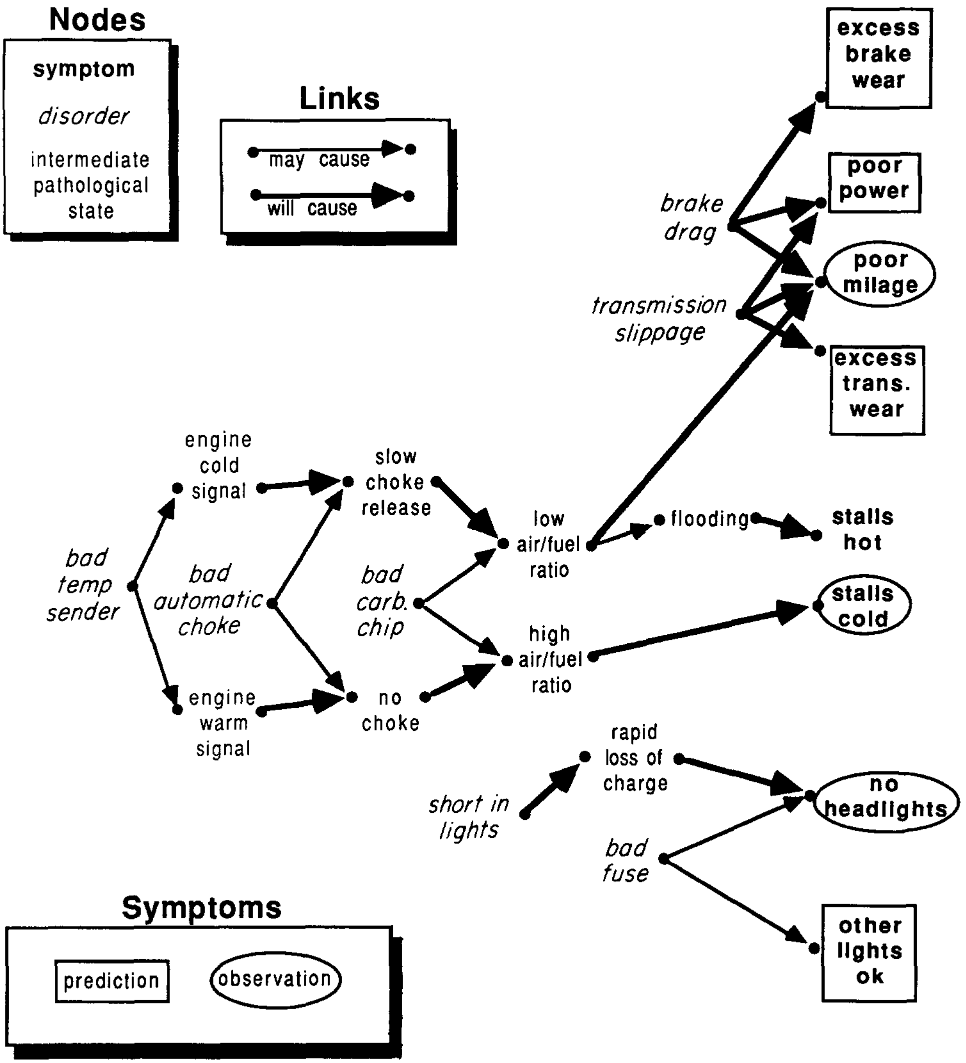

Fig. 2 A patient-specific model (PSM) for the three symptoms.

the PSMs to take the observed finding into account. Finally, this process is repeated starting at Step 2.

The algorithm in Step 2 for deciding which PSMs will be expanded into Diagnostic Closures is very simple. The PSM with the smallest total of unaccounted or partially unaccounted states is deemed most promising. Once that is decided, the algorithm for deciding which hypothesis (set of disorders within that PSM) will be explored considers *compatibility* and *testability*. If a predicted manifestation of a hypothesized disorder seems incompatible with the reported symptoms, that lowers the score of that hypothesis. If a predicted manifestations is compatible, and is easily tested, and is useful for differential diagnosis, that increases the score of that hypothesis. Note that the hypothesis scoring mechanism measures internal consistency and the usefulness of confirmatory evidence, not likelihood; if a rare disorder meets these criteria, it will become the focus of the information-gathering phase.

To continue the example of Fig. 2, note that according to the causal model, none of these three symptoms (bold-face nodes) may cause any of the others. So, in this simple example, there is only a single initial PSM.[3] The Diagnostic Closure consists of the projected manifestations of the reported symptoms (none in this example), all the potential causes of the unaccounted symptoms, and all the projected manifestations of the potential causes (boxed nodes). In forming plausible hypotheses from this Diagnostic Closure, the incompatibility of *low air/fuel ratio* with *high air/fuel ratio* will eliminate any hypothesis that explains *stalls when cold* by *no choke action* but explains *poor mileage* with *slow choke release*. *ABEL* in this case gives top score to the hypothesis which explains *poor mileage by dragging brake*, *stalls when cold* by *bad choke*, and *no headlights* by *bad fuse* (not shown). Its high score is based on its internal consistency and the fact that three projected manifestations of these causes — *poor power*, *excess brake wear*, and *other lights OK* — are easily tested. After these questions have been answered by the user, the new observations are added to the PSM, alternative PSMs are generated as needed, and the cycle repeats. *ABEL*'s final product will be the highest-ranked multiple-disorder hypothesis.

Parsimonious set covering — a mathematical approach

A long-term research effort by Reggia *et. al.*, has developed the Parsimonious (or General) Set Covering model (GSC) of diagnosis (1985a and b). This model was motivated in part by the feeling that it was incorrect to deal with each disorder individually in a multiple-fault diagnosis, as *INTERNIST* does. Instead of probabilities, GSC employs a causal relation which records all possible manifesttions of each disorder and all possible causes of each symptom.[4] A diagnostic problem is defined as $P = (\mathbf{D.M.C.M^+})$ where $\mathbf{D}$ is the set of all disorders, $\mathbf{M}$ is the set of all manifestations, and $\mathbf{C}$ is the causal relation. A pair $\langle d_i m_j \rangle \in \mathbf{C}$, where $d_i \in \mathbf{D}$ and $m_j \in \mathbf{M}$ means 'disorder *i* may cause manifestation *j*'. M^+ is the subset of $\mathbf{M}$ which contains all symptoms reported.

In this setting, the term *manifs(d_i)* means the set of all manifestations of disorder d_i and the term *causes(m_j)* means the set of all disorders which can cause m_j. An *explanation* E^+ for M^+ is a set of disorders where $M^+ \subset$ *manifs(E^+)*and E^+ is *parsimonious* in some sense. The principle of parsimony is, in effect, Occam's Razor. During most of the research into GSC theory, parsimony has been interpreted as 'minimum cardinality'.

In this model, *causes(M^+)* is the universe of all possible disorders which could cause at least one of the manifestations in M^+. The diagnostic task is viewed as choosing a small number of candidates for E^+ and then performing differential diagnosis in the traditional sense — identifying a few easily-obtained symptoms which can rule out all but one of the candidates. Note that a candidate for E^+ is a set of disorders which together explain all the symptoms, so a candidate is inherently a multiple-fault hypothesis.[5]

Table 7 reflects the example knowledge base as GSC would represent it via the Causal Relation. According to this Table, *manifs*(bad timing) is the set {poor mileage, poor power}, and *causes*(stalls when cold) is the set {bad carburettor chip, bad auto choke, bad temperature sender}.

Table 7. A GSC-style causal relation for the auto repair knowledge base

Disorder	*Manifestation*
Bad alternator	Dead battery
Bad auto choke	Poor mileage, stalls when cold
Bad battery	Dead battery
Bad carburettor chip	Poor mileage, stalls when cold
Bad carburettor chip	Poor power, stalls when hot
Bad fuse	No headlights
Bad temperature sender	Stalls when cold
Bad thermostat	Overheating
Bad timing	Poor mileage, poor power
Bad voltage regulator	Dead battery
Clogged radiator	Overheating
Dragging brake	Poor mileage, poor power
Gas line leak	Poor mileage
Major short	Dead battery
Short in lights	No headlights
Trans/clutch slippage	Poor mileage, poor power

The diagnostic algorithm uses three sets. *MANIFS*[6] is the set of all manifestations reported so far. *SCOPE* is *causes(MANIFS)*, and *HYPOTHESIS* is the set of all combinations of disorders which could explain *MANIFS*. The basic cycle is:

1. Get the next manifestation m_j and add it to *MANIFS*.
2. Retrieve *causes*(m_j) from the Causal Relation.
3. Add (*causes*(m_j) to *SCOPE*.
4. Adjust *HYPOTHESIS* to accommodate m_j.
5. Repeat until no further manifestations are reported.

An important element of the GSC model is the compact representation of *HYPOTHESIS* as a set of *generators*(Reggia *et. al.*, 1985b). Instead of explicitly recording each combination of disorders that could account for (or 'cover') M^+, GSC records the sets of terms whose Cartesian product is the set of all combinations in *HYPOTHESIS*. For example, the set $G = \{\langle d_1,d_2\rangle, \langle d_3,d_4\rangle, \langle d_5,d_6\rangle)$ would generate the set of disorder triples $\langle d_1,d_3,d_5\rangle$, $\langle d_2,d_3,d_5\rangle$, $\langle d_1,d_4,d_5\rangle$, $\langle d_2,d_4,d_5\rangle$, $\langle d_1,d_3,d_6\rangle$, $\langle d_2,d_3,d_6\rangle$, $\langle d_1,d_4,d_6\rangle$, and $\langle {}_2,d_4,d_6\rangle$.

In addition to its compactness, the generator representation parallels the manner in which human clinicians think and talk about a multiple-fault hypotheses. Using the previous example, description (i) is preferred over (ii):

(i) The patient has either d_1 or d_2 together with either d_3 or d_4.
(ii) The patient has either d_1 and d_3 or d_2 and d_3 or d_1 and d_4 or d_2 and d_4.

It also makes possible an efficient way of manipulating these sets, using *generator division* and other special operations. Step 4 of the algorithm above uses generator division, dividing the old *HYPOTHESIS* by the set *manifs*(m_j) to produce a new *HYPOTHESIS which accounts for the newly reported symptom* m_j.

In the automobile example, GSC would start with a blank slate and ask for the initial manifestation. Assume that *poor mileage* is first mentioned, and thus is the sole inhabitant of *MANIFS*. *SCOPE* would then be the set (bad auto choke, bad carburettor chip, bad timing, gas line leak, dragging brake, trans/clutch slippage). *HYPOTHESIS* would be $|\langle bac,bcc,bt,gll,db,ts\rangle|$;[7] that is, a single hypothesis which is represented as the generator 'one of bac, bcc ...'.

In the second round, *stalls when cold* is volunteered, and is added to *MANIFS*. *SCOPE* has *bad temperature sender* added, and becomes (bac,bcc,bt,bts,gll, db,ts). *HYPOTHESIS* is updated by generator division-with-remainder[8] as follows:

$$\langle bac,bcc,bt,gll,db,ts\rangle \div \langle bac,bcc,bts\rangle$$
$$= \begin{bmatrix} \langle bac,bcc\rangle \\ \text{and} \\ \langle \text{bt,gll,db,ts}\rangle \times \langle \text{bts}\rangle \end{bmatrix} \qquad (2).$$

The hypothesis represented by this generator is read as '**either** one of ⟨bac,bcc⟩, **or** one of ⟨bt,gll,db,ts⟩ combined with bts', which agrees with intuition.

In the third round, when *no headlights* is reported, *SCOPE* is augmented by (bad fuse, short in lights) and *HYPOTHESIS* becomes

$$\begin{bmatrix} \langle bac,bcc\rangle \times \langle bf,sil\rangle \\ \text{and} \\ \langle \text{bt,gll,db,ts}\rangle \times \langle \text{bts}\rangle \times \langle bf,sil\rangle \end{bmatrix} \qquad (3).$$

and so on. When the user ceases to volunteer information, *HYPOTHESIS* is analysed to identify a differential diagnosis problem which is parsimonious.[9] A manifestation is chosen which is not in *MANIFS* but is in *manifs*(d) for some disorder d in the hypothesis, and a question generated concerning that symptom. In the example, the minimum cardinality hypothesis will contain either *bad choke* or *bad carburettor* combined with either *bad fuse* or *short in lights*. Therefore, questions will be chosen to discriminate among those pairs of contenders. GSC's final product will be a single multiple-disorder hypothesis which covers all reported symptoms and is as simple (parsimonious) as possible.

Diagnosis from first principles — an approach based on first-order logic

A number of researchers have attempted to formalize the diagnostic process in first-order logic (Genesereth, 1984; Reiter, 1985; deKleer & Williams, 1986b). We will take Reiter's work as characteristic of this approach, although significant differences exist among them.

We could describe any system — the human body or an engine — as composed of COMPONENTS, a finite set of parts, sub-systems, etc., and SD, a set of first-order sentences which axiomatize a system description for normal operation. SD contains a distinguished unary predicate AB, which means 'abnormal'. Thus, for a model of the electrical system in our example, COMPONENTS might be {battery, cable-pos, cable-neg, starter, ...} and SD might contain such axioms as:

$$\text{Auto}(x) \wedge \text{Battery-Of}(x,b) \wedge \neg\text{AB}(b) \Rightarrow$$
$$\text{voltage}(b) \geq 10 \wedge \text{voltage}(b) \leq 19 \qquad (4).$$

A diagnostic problem is a properly defined system combined with a set of observations, OBS, which describe the actual behaviour of the system, or symptoms, such that

$$\text{SD} \cup \text{OBS} \cup [\forall c \in \text{COMPONENTS}\ \neg\text{AB}(c)] \qquad (5).$$

is inconsistent. A diagnosis for this problem is a minimal set F of faulty compnents; i.e. if we assume each member of F is abnormal and all other members of COMPONENTS are normal, we achieve consistency. However, computing this directly could be undecidable or intractable. So two other types of sets are introduced: *conflict sets*, due to deKleer, and *hitting sets*.

A conflict set CS is a collection of components such that

$$\text{SD} \cup \text{OBS} \cup [\forall c \in \text{CS}\ \neg\text{AB}(c)] \qquad (6).$$

is inconsistent. A *minimal conflict set* has no proper subset which is a conflict set. Intuitively, a conflict set is a conflict set in which every member participates in at least one possible fault. Unless the symptoms are so conclusive that only one combination of disorders could possibly be correct, there will be many minimal conflict sets for a diagnostic problem. If S is a collection of minimal conflict sets, then a *hitting set* for S is a set H which has a non-null intersection with each (minimal conflict) set in S.

For example: if OBS was a logical representation of 'the gas mileage is poor' and 'it stalls when cold', then one minimal conflict set would be {*auto choke, temperature sender, carburettor chip*}. Another minimal conflict set would be {*auto choke, carburettor chip, timing, gas line, brakes, transmission*}. Intuitively, each of these is a conflict set because it represents all the components whose failure could cause one of the symptoms — if we postulate $\neg AB(c)$ for each member of the set, that is inconsistent with the observation that *something* must have caused the symptom.

If the set of all conflict sets, S, consisted of just the two mentioned above, then the set of minimal hitting sets, HS, would contain {*auto choke*}, {*carburettor chip*}, {*temperature sender, timing*}, {*temperature sender, gas line*}, {*temperature sender, transmission*} and {*temperature sender, brakes*}.

The end result is that a set $\Delta \subset$ COMPONENTS is a diagnosis **iff** Δ is a minimal hitting set for the set of conflict sets in the diagnostic problem.

While acknowledging that, in general, the consistency of arbitrary collections of first order formulae is undecidable, Reiter claims that most diagnostic expert systems operate in arenas where a general or special-purpose theorem prover can be used successfully to test the consistency of subsets of a diagnostic problem. He offers an 'algorithm' for computing the set of all diagnoses; it is expected to be computationally tractable in most real-world situations. It requires computing the set of all minimal conflict sets for the problem (or at least the first n minimal conflict sets) and building a special structure called an *HS-tree* (for Hitting Set tree).

In those cases where consistency is computable, a theorem prover which builds refutation proofs of inconsistency may be used to generate counter-examples for

$$\text{SD} \cup \text{OBS} \cup [\forall c \in \text{COMPONENTS} \neg\text{AB}(c). \qquad (7).$$

In each counter-example, the components dealing with AB will define a conflict set. Thus, Reiter presumes that a function TP (implemented as some kind of theorem prover) can be defined in any given arena which will return a stream of conflict sets.

The function TP can be used to build an HS-tree, as follows (see Fig.3):

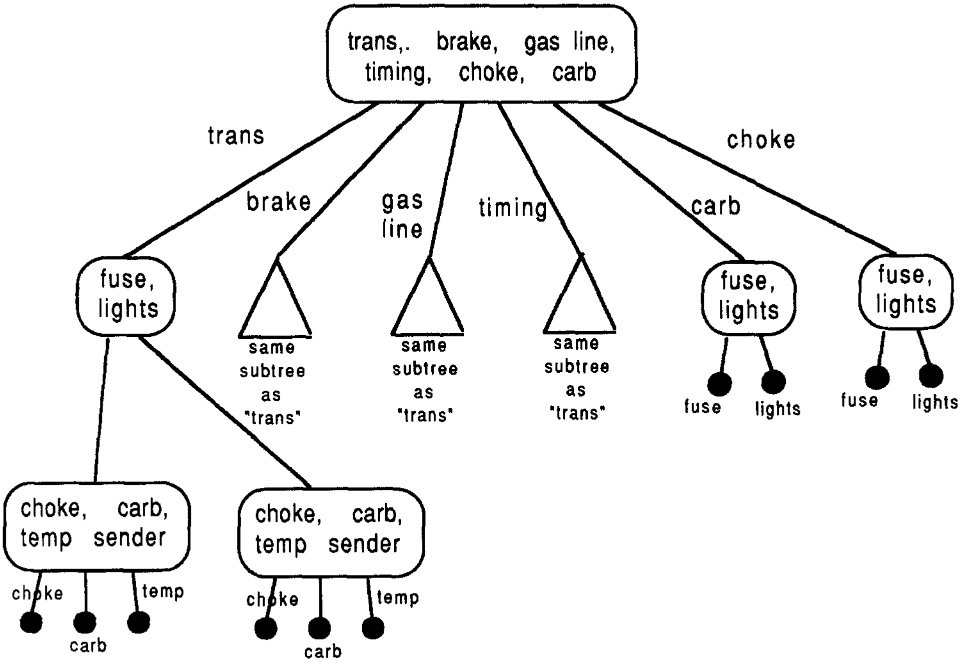

Fig. 3 An HS-tree for the auto repair example, assuming *poor mileage, stalls when cold* and *no headlights*.

1 Label the root node (node 0) with any arbitrary conflict set returned by TP(SD, OBS, COMPONENTS).
2 For each element in the set which labels this node, create a child node and label the arc to the child node with that element.

3 For each unlabelled node, do:

(a) Let $H(n)$ be the set of elements found on arc labels on the path from this node n to the root.

(b) If any other node in the tree has a label L such that $L \cap H(n) = f$, then label this node with L. If not, then call TP(SD, OBS, COMPONENTS − $H(n)$; if it returns a non-null conflict set, label this node with that set. Otherwise, label this node with f.

(c) If this node n is labelled with f, and there is another node in the tree v such that $H(n) \subseteq H(v)$, then mark node v 'closed'. Do not create a label or children of v.

(d) If this node is not labelled with f then create child nodes as in Step 2 above.

4. If all leaf nodes are labelled with f, then terminate tree-building. Otherwise, repeat the above step.

When the HS-tree is finished, the set of all diagnoses is $\{H(n) \mid n \text{ is labelled with } f\}$, i.e. the set of all minimal hitting sets for conflict sets of the problem. Note that the path to each leaf node in Fig. 3 defines a minimal hitting set which belongs to the set of potential diagnoses for our example problem.

3 Comparison of these approaches

In this section we discuss the strengths and weaknesses of these five approaches in handling multiple fault diagnoses. We will examine in particular their Problem Formulation mechanisms — the manner in which each system constructs a manageably small sub-problem from a potentially enormous search space. We recognize that each of the implemented systems (*INTERNIST*, GSC (many times), and *MEDAS*) was built to operate in a very narrow but real domain. That domain's influence on system design cannot always be separated fully from the theoretical underpinnings, but we shall try.

Relaxing the Bayesian assumptions

The standard version of Bayes Theorem requires three assumptions: (1) the set of disorders must be mutually exclusive and exhaustive; (2) the set of symptoms must be independent of one another in general; and (3) the set of symptoms associated with a particular disorder must be independent given the presence of that disorder. A practical diagnostic system must relax these assumptions in some way in order to reduce the statistical data required and to simplify the computations involved. The two Baysean systems we have discussed, *MEDAS* and *INTERNIST*, differ in how they address the problem.

MEDAS takes the first assumption very seriously, and treats a diagnostic problem as a series of binary choices: 'is disorder x present or nor?' By restricting the probabilistic decision to x or ¬x, it meets the first condition, but at a price. The Binary Choice strategy always looks at a collection of small decisions, never at the whole diagnostic problem.

This can lead to inefficiency. The *MEDAS* diagnostic cycle is driven by the goal 'if any disorder has partial supportive evidence, and little disconfirming evidence, seek more evidence until the disorder's existence is either reasonably established or substantially disconfirmed'. This is not a differential diagnosis approach — there is no attempt to find the evidence which can best decide between two likely disorders. If many alternative explanations for the same set of symptoms are possible (i.e. a family of disorders) the system may devote effort to strengthening the already-strong hypothesis that 'one member of this family is present' instead of focusing on the differential question 'which member of the family is it?' The problem of differential diagnosis — which many feel is the crux of the problem (Pople, 1982, p.120) — is thus off-loaded largely to the physician.

Partially to compensate for this restriction, *MEDAS* relies heavily upon the user/physician for (heuristic) guidance. At the end of each round, the system displays a list of the top ten symptoms or tests and asks the physician which one to pursue. The (ranked) list of disorders under consideration is frequently displayed, but the physician is asked to prune or supplement it based upon professional judgement (Ben-Bassat, et. *al.*, 1980, p.154). To the extent that the list of ranked disorders rapidly provides useful information to the physician, *MEDAS* achieves its goal. But we would argue that to the extent the physician is influenced by the comparison of ranks, *MEDAS* departs from pure binary choice. If the physician is invited, implicitly, to compare the computed probabilities, then the system has allowed a violation of Bayesian assumption no.1 without taking advantage, in its internal calculations, of the removal of that restraint.

Although Binary Choice does deal with multiple disorders, it does not deal directly with formulation and resolution of differential diagnosis problems. This seems a critical deficiency, in general. The *MEDAS* system is not, however, offered as a general approach to machine-directed abductive reasoning problems, but as a system designed for swift, large-grain decision-making in an emergency room. *MEDAS* does seem well suited for this class of applications.

MEDAS adheres to Bayesian Assumption no.1, while readily admitting that the second and third assumptions are violated 'in most realistic cases'. Ben-Bassat argues that the practical difficulties in obtaining estimates of the conditional probabilities for clusters of symptoms are considerable — and would introduce more error into the system than allowing the violations to go uncorrected (Ben-Bassat et. *al.*, 1980, p.153).

INTERNIST deals with Assumption no.1 by using an alternate form of Bayes' Theorem (see equation (1), but likewise accepts wholesale violation of the other two assumptions. The practical necessity of assuming (unrealistic) statistical independence of symptoms caused some to reject Bayesian statistics as a basis for abductive reasoning systems. But Charniak (Charniak, 1983) has pointed out that violation of Assumption no.2 introduces an error in the denominator of equation (1). This affects the *absolute* magnitudes of the probabilities calculated, but not the *relative* magnitudes. Thus, in any differential diagnosis problem, where magnitudes of competing disorders are being compared, the comparisons are unaffected by violation of the second assumption.

Charniak also points out that violations of Assumption no.3 can be dealt with inside the Bayesian framework. In the automobile example (see Table 1), assume a short in the lighting circuit — if it is severe — can disable the entire electrical system, bringing about a large collection of symptoms. A mechanic who suspects a severe short will check one of those symptoms, perhaps the radio, confident that all or none of the symptoms will appear. This completely violates Assumption no.3. The mechanic reasons about the situation via intermediate pathological states, or syndromes. By introducing such an intermediate state (e.g. system short out) explicitly and computing the conditional probability of symptoms given the present of the intermediate state, statistical soundness is restored. Only two assumptions are required for this approach to hold.

1 The symptoms (e.g. no horn, no radio) are related to the ultimate disorder (short in lights) only via the intermediate state (system short out).
2 The symptom and the ultimate disorder must be independent given the presence of the intermediate state.

These are far less stringent restrictions, and are reasonable in many cases (Charniak, 1983).

Thus the criticisms of the Bayesian approach are largely without effect.[10] They do not impeach its usefulness in identifying the most probable disorder hypotheses, which is a central step in the formulation of differential diagnosis problems. Of course, the Bayesian approach does not offer any direct assistance in finding the correct (or most reasonable) combination of disorders in a multiple-fault diagnosis. In models such as *INTERNIST*, heuristics take control of the search once a manageably small problem (or series of problems) has been formulated.

Reasoning about intermediate states

Constructing an accurate, rich causal model of a diagnostic domain requires the use of intermediate states which lie between the root hypotheses and the symptoms. This is obviously crucial for generating good explanations (Swartout, 1083), but it also has a major impact on the diagnostic process itself.

Patil's primary goal with *ABEL* was to investigate the power of a deep causal model for guiding abductive reasoning (Patil, 1981). On purely statistical grounds, Charniak argued that intermediate pathological states must be explicitly treated (Charniak, 1983). Pople originally rejected intermediate states and causal models for *INTERNIST-I*, based primarily on the poor performance of the earliest *INTERNIST* version, which employed such a model (Pople, 1977, p.1031). He still argues that a causal model *per se* is insufficient, but now believes a robust system should incorporate both a causal model and an *n*-dimensional nosological hierarchy, with planning links between them (Pople, 1982, p.161). Reggia *et. al.* enhanced their bi-partite model of GSC early on to accommodate intermediate states (Reggia & Peng, 1986, p.21). And all efforts under the banner of 'diagnosis from first principles' rely upon a complete model of the processes going on in the system being diagnosed.

It seems clear that there is a set of core arguments in support of causal reasoning in adbuctive systems.

(i) When manifestations are linked directly to ultimate causes (disorders), we discard information about the mechanisms by which those disorders produce the manifestations. Discarded information cannot be exploited.

(ii) Many intermediate states (syndromes) are easily-recognized clinical phenomena. Some are associated with pathognomonic symptoms, and testing for them can quickly narrow the scope of the differential diagnostic problem.

(iii) Causal reasoning seems to mimic experienced clinicians, who use whatever is known about the suspected disease process to test their diagnostic hypotheses (Kassirer & Gorry, 1978)

Yet despite this consensus on the value of causal models, there are practical difficulties. The primary difficulty is that a rich causal model is much more difficult to construct. The power of causal models is limited by the depth and accuracy of the knowledge they contain. That varies greatly from domain to domain and even within a well-studied field like medicine. In particular, the diagnosis of artifacts such as electronic circuits offers greater opportunities for causal modelling (deKleer & Williams, 1986a and b). In a Bayesian-based system, one way the complexity of introducing intermediate states manifests itself is in the many-fold increase in the number of prior probabilities to be gathered or estimated.

Intermediate pathological states provide local foci for differential problem formulation and 'milestones' in the diagnostic process. But the introduction of many small, intermediate steps may prevent the large logical leaps which expert clinicians exhibit. To avoid that loss requires more sophisticated control strategies.

A fully implemented *CADUCEUS*, as the evolving system is now called, may provide more insights into the costs and benefits of exploiting causal models. Hopefully, it will also explore the control issues involved in integrating causal reasoning with nosological, statistical and other forms of knowledge. We must await clinical evaluation of *CADUCEUS*, however, to learn how much of a performance improvement can be attributed to causal reasoning.

Sequential sub-problems vs. multiple fault hypotheses

A major difference between the Bayesian approaches and the others is that conditional probability offers no mechanism for developing and evaluating multiple-fault hypotheses. Bayesian systems rely upon heuristics for that. *INTERNIST*-I groups symptoms according to major organ involvement (electrical problems or fuel system problems, in our example) and then forms differential diagnosis subproblems (by organ group) one at a time. The heuristic is to identify the most probable single disorder and form a differential problem around it, including all other disorders which might explain essentially the same symptoms as the top-ranked candidate. After identifying a clear winner, the process is repeated for another group of symptoms, but now 'bonus points' are awarded to any hypothesis that is associated with a previous winner.

The major drawback of this sequential sub-problem approach is that the *interaction* of disorders cannot be fully considered. The 'bonus point' scheme alleviates this somewhat, after some of the diagnostic sub-problems have been solved. But this approach misses most of the power available from reasoning about disorder interaction.

Another disadvantage of this heuristic is that sometimes a symptom could contribute to the probabilities of disorders in several different differential groups. In the automobile example, warm stalling could be a symptom of fuel system problems (bad choke or bad carburettor chip) or, in rare cases, a symptom of very low alternator output or a short just severe enough to cancel out the alternator's output. The sequential sub-problem approach, if it focused first on the electrical problems, might explain warm stalling as an electrical manifestation, and thus deprive the fuel system sub-problem of a valuable clue. We repeat that this is not a failure of the Bayesian approach, but a criticism of the heuristics used in *INTERNIST*-1 to extend it to multiple-fault situations.

The other approaches, by contrast, construct multiple disorder hypotheses which account for all the symptoms observed to date. This enables them to reason about the interactions of those disorders encompassed by the hypothesis[11] and neatly avoids the question of which disorder 'owns' any one symptom. Formation of the multiple-disorder hypothesis is grounded in the underlying theory of that approach; it is not heuristic.

This is clearly desirable, but immediately gives rise to some difficulties. Potentially, *ABEL* and GSC both must deal with the combinatorics that *INTERNIST*'s heuristic avoids. In fact, GSC's use of generators to represent the sets of possible diagnoses is motivated in part by the large numbers of combinations that may be candidates at any one time. So, heuristics are still used to select the hypotheses to be explored in each round of their diagnostic cycles. In basic GSC and approaches based on first-order logic, the number of disorders contained in a hypothesis is used — the principle of parsimony is invoked. In *ABEL*, the number (and perhaps magnitude) of 'loose ends'[12] is used to identify the best multiple-disorder hypothesis.

Because GSC uses the principle of parsimony for selecting hypotheses, it encounters the 'Rare Disease Problem'. Assume that there is a rare engine disorder in our example which causes poor mileage, poor power, cold stalling, and battery failure. By the principle of parsimony, that rare disorder should be invoked whenever all those symptoms are present. It is a parsimonious diagnosis, but not a very useful one. Similarly, in *ABEL* it is possible that this rare engine disorder might seem to tie up all the loose ends, and it could be chosen for exploration, despite its improbability.

The point is this: all of these approaches go as far as their underlying theory will take them, and then use heuristics to carry on with differential diagnosis. Those heuristics have shortcomings, as outlined here. We believe it is clearly better to have a sound theoretical basis for formulating the multiple-disorder hypotheses. The ideal would be to go one more step — to evaluate the alternative multiple-disorder hypotheses in a theoretically sound way. A recent extension of GSC would seem to accomplish this; it is discussed in Section 4 below.

The meaning of parsimony

Almost all diagnostic approaches invoke some form of *Occam's Razor* or the *principle of parsimony* to choose between competitive hypotheses. They differ, however, in how parsimony is defined.

In the GSC approach, a diagnosis has been defined as a parsimonious cover for the set of manifestations M^+. Parsimony, in turn, has been defined as 'containing the smallest number of disorders'. This definition has been implemented in a number of prototype systems, and performed reasonably well in real-world settings. But it is vulnerable to the 'Rare Disease Problem'. Reiter (Reiter, 1985) and deKleer (deKleer & Williams, 1986a) employ a less severe form of Occam's Razor and view parsimony as irredundancy. In their approaches, a diagnosis is defined as 'the set of *irredundant* covers of M^+'. An *irredundant cover* is one that has no proper subset which is a cover. It is interesting that Reggia *et. al.* reached the same conclusion concurrently, but independently, by a different line of reasoning (Reggia & Peng, 1986, p.22).

Quantified symptoms

Most diagnostic systems view symptoms as propositions. They are either present or absent or, perhaps, present to some degree. Patil's *ABEL* has the unique feature of quantified symptoms and provision for quantitative relationships within the causal model. For example, the causal relations of the other approaches contain information like 'a bad automatic choke may cause cold stalling'. *ABEL*'s causal model can contain information like 'a bad automatic choke may cause the air/fuel ratio not to be elevated by 20% when the COLD signal is ON' and 'not elevating the air/fuel ratio 15–25% when the engine is cold may cause poor starting or stalling'. The claimed benefit of causal model reasoning is that knowledge about the intermediate states (e.g. air/fuel ratio) can generate predictions about the manifestations of those intermediate states, and those predictions are a valuable source of confirmatory evidence. But we believe an additional source of power in *ABEL*'s approach is the quantitative nature of some of that reasoning, e.g. 'elevated by 20%'. This enables efficient 'netting out' of the effects of arbitrarily many complementary or offsetting causes. It also allows for 'sharing' of a symptom by two unrelated disorders and for reasoning about the unique combined effects of some disorders.

Note that this quantitative reasoning is simulated to some extent in those systems that distinguish among *trace, mild, moderate, strong* and *severe* levels of a symptom or intermediate state (e.g. x may cause mild hypertension). But combining or netting effects is more difficult and less precise when qualitative descriptors are used.

Causal relations, etc. could be structures to accommodate this. For instance, *INTERNIST*'s *MANIFESTS* relation (see Table 5 in Section 2), if it recognized intermediate states, could record for symptom = 'air/fuel ratio too low when engine cold' a strength = 4 and an *amount* = 17%. Thus, quantitative reasoning is not limited to systems employing causal models, and it need not always involve intermediate states — it could quantify the symptom of a disorder. However, the particular domain for which *ABEL* is designed lends itself to quantification; most symptoms are laboratory measurements of ratios, and most knowledge about causation is in terms of increases or decreases from stable, normal levels. Other domains may not lend themselves to such precision.

Equivalence of formalisms

Reiter has shown (Reiter, 1985, p.33) that the GSC formulation of a diagnostic problem can be transformed in a straightforward manner to his formalism. Of course, the GSC formalism is more restricted than full first order logic, with less expressive power. To see how to transform a diagnostic problem in Reiter's formalism into an equivalent GSC formulation we first note that the sentences in the System Description SD are of two types: *normative* descriptions of operation, such as

$$\text{battery(b)} \wedge \neg \text{AB(b)} \Rightarrow \text{voltage(b)} \leq 14 \wedge \text{voltage(b)} \geq 9 \qquad (8)$$

and descriptions of *causation* such as

$$observed(m) \Rightarrow present(d_1) \vee \ldots \vee present(d_j). \qquad (9).$$

The normal operating values are always either a single value, a range of values, or a (disjunction of) predicate(s).

A procedure for transforming such a first order diagnostic problem into GSC form would obviously convert all sentences of the *causation* type directly into pairs in the Causal Relation, i.e. $\langle d_1, m..., \cdots, \langle d_j, m \rangle$. Sentences of the *normative* type would first be turned into converse, e.g.

$$\neg \text{voltage(b)} \leq 14 \vee \neg \text{voltage(b)} \geq 9 \Rightarrow \neg \text{battery(b)} \vee \text{AB(b)}. \qquad (10).$$

Then each term on the left hand side of the converse is translated (if necessary) into the proper terminology for symptom description (e.g. low-voltage). Finally, each combination of left hand side and right hand side terms produces a pair in the Causal Relation. The rest of the transformation is straightforward.

So long as parsimony is defined as irredundancy in the GSC problem, equivalent results will be achieved from either formulation of a problem. It is difficult to say which formalism might be the more tractable. The algorithm used in GSC is nearly exponential in the worst case, but Reggia *et. al.* report that in all problem domains studied so far, computational efficiency has not been a problem. Reiter's 'algorithm' depends upon a theorem prover component proving the consistency of a collection of first order sentences, which in general is undecidable. By that analysis, GSC is the lesser of two evils. The expected performance of each algorithm depends on such things as the average size of *causes*(M^+) in GSC or the average cost of a refutation proof in Reiter's formalism. This would seem to be a fruitful area for further research.

4 Summary and an emerging consensus

We will begin this section by trying to summarize the issues discussed in comparing the approaches in the previous section. This will lead to an emerging consensus for a formalization of the diagnostic process.

Summarizing the comparisons

The Binary Choice strategy of Ben-Bassat *et. al.* is one way of living with the strict assumptions of statistical independence required with Bayesian statistics. However, there are other ways to deal with this (e.g. *INTERNIST*'s alternate formulation of Bayes' Theorem) which preserve the ability to compare posterior probabilities of different disorders. Therefore, the Binary Choice strategy is not a good general model for diagnostic systems.

All of the independence assumptions required in Bayesian statistics are unrealistic in most diagnostic settings. This has led some to reject all Bayesian approaches as unsound. However, Charniak has successfully answered most of the charges levelled against the Bayesian model, showing that:

(i) The requirement that the set of disorders is mutually exclusive and exhaustive can be avoided by using an alternative formulation of Bayes' Theorem.

(ii) The requirement that all symptoms be independent can be safely ignored so long as the *absolute* magnitudes of the posterior probabilities are not used, only the *relative* magnitudes.

(iii) The requirement that all symptoms of a particular disorder be independent given the presence of that disorder can be dealt with successfully via intermediate pathological states.

Nearly all the researchers agree that intermediate states are useful for encoding knowledge about how a disorder causes a manifestation, and as focal points early in diagnosis, especially when they have unique pathognomonic cymptoms. There are drawbacks:

(a) there is insufficient understanding, in some domains, to build an accurate causal model,

(b) they require gathering or estimation of even more statistics in Bayesian systems,

(c) they introduce so much detail that they sometimes can make it hard for the diagnostic system to 'see the forest for all the trees'; thus they often require more complex control strategies.

It remains to be seen whether the incorporation of detailed causal reasoning in an otherwise complete system will provide a performance improvement commensurate with the additional costs. Field trials of *CADUCEUS* seem the best hope for gathering solid evidence on this issue.

The Bayesian approach provides a sound mechanism for identifying likely disorders to explain clusters of symptoms, but no mechanism to formulate a multiple-disorder hypothesis which explains all the observed symptoms. *INTERNIST* has a very workable heuristic for that purpose. The GSC and formal logic approaches have a sound mechanism for formulating multiple-disorder hypotheses, but not for choosing one of those to pursue. The latter is a better position to be in, simply because it allows us to get further along in the diagnostic process before resorting to heuristics. A recent development which takes us further still is discussed below.

Although the GSC model has long used *minimum cardinality* as the definition of parsimony — and hence as the criterion for a 'best diagnosis' — both Reggia *et. al.* and Reiter have concluded that *irredundancy* is the proper meaning of parsimony.

ABEL uses quantified symptoms, and a quantified causal model, to considerable advantage. This facilitates several desirable traits for diagnostic systems:

(a) a symptom which is partially explained by several disorders can be 'allocated' mathematically among them;
(b) a symptom can be viewed as fully or partially masked by the offsetting effects of several disorders;
(c) combination or offsetting of effects can be done more precisely than in a qualitative system.

Quantified symptoms are not restricted to detailed causal models, although they fit well with that approach.

Reiter has shown that the GSC formulation of a diagnostic problem can be transformed easily into his own first-order logic formalism. We show that with one constraining assumption, the reverse transformation can be done also. It is not clear whether either approach is computationally more tractable in all cases; both have worst case costs that are intimidating.

On a more pragmatic note, Ramsay *et. al.* (Ramsay, *et. al.* 1986) surveyed the results of six comparisons, conducted between 1971 and 1980, of Bayesian, GSC, and rule-based diagnostic systems — all in medical settings. They concluded that no method was significantly superior to the others in accuracy, relative ease of use and ease of construction. The differences they did encounter were attributed to differences in the problem-specific information in the knowledge bases as opposed to fundamental differences in the methods being used (Ramsay *et. al.*, 1986, p.482).

An emerging consensus

It was pointed out above that GSC can be transformed into Reiter's first-order logical formalism and *vice versa*. From here on we will treat them as equivalent and call them GSC. It was also pointed out that, in theory at least, all the researchers agree on the value of intermediate pathological states and hence, causal models. Indeed, except for the fundamental difference between the Bayesian and GSC models, it could be said that the *CADUCEUS* design represents a consensus on the various types of reasoning needed for truly expert performance in a diagnostic system:

(i) Bayesian posterior probabilities freed of most restrictions regarding independence;

(ii) causal reasoning based upon a detailed model;
(iii) 'taxonomic reasoning' based upon numerous nosological models.[13]

Some very recent results appear to bridge that fundamental difference, making a complete consensus possible.

Peng & Reggia (1987a) report a method for developing a set of parsimonious covers for the symptoms (using the GSC formalism) and guaranteeing that the probability of the correct diagnosis being in the set is not less than any *Comfort Measure* (CM) we wish to establish. Their method does not require any stronger assumptions of statistical independence than *INTERNIST*, and it uses only the prior probabilities of each disorder and the *causal strength*[14] of each disorder for each of its manifestations. Furthermore, it preserves a measure of the relative likelihood of each cover in the set, so that the most probable (multiple disorder) diagnosis is automatically identified.

The mathematical details are cumbersome but they propose an algorithm which exploits these key ideas:

1 It is tractable to calculate a relative likelihood measure $L(D_I \mid M+)$ for any hypothesis D_I; it differs from the posterior probability of D_I by a constant factor. Thus, the largest L always signifies the most probable hypothesis.
2 It is tractable to calculate $UB(D_I,M^+)$, an upper bound on the relative likelihood of any proper superset of D_I.
3 The *sum* of the relative likelihoods (L-values) of all proper supersets of any D_I can be calculated.
4 Whenever some hypothesis D_k which is a cover for $M+$ has the K^{th} largest L-value of all cover hypotheses, and its L-value exceeds the UB-value of any non-cover hypothesis which has been generated to date by the algorithm, then D_k is the K^{th} most probable hypothesis among *all possible* hypotheses.

Although this method has not been implemented or tested to our knowledge, it could make possible a diagnostic paradigm which meets the consensus desiderata of all the researchers surveyed here. That assumes, of course, that this Bayesian GSC style of problem formulation can effectively complement nosological and causal reasoning, and that a control strategy can be devised which draws upon the strengths of each reasoning style at the appropriate time in a calculate–hypothesize–question cycle. The creation of such a consensus system in a complex, real-world domain would be significant effort, but potentially a very rewarding one.

[1] see Fig. 1. The lower bound of probability ranges is used throughout this example.

[2] Only the *EVOKES* and *MANIFESTS* relations are shown in our example. Other relations are defined on the set of disease entities to record the causal, temporal, and other relations between diseases. *INTERNIST* would also show, for instance, the association between bad carburettor chips and bad chokes which is recorded in Table 2.

[3] In a more realistic problem, there are typically numerous alternative interpretations of the reported data. Each gives rise to a separate PSM.

[4] The necessity of intermediate pathological states was recognized early in this research, and they were included in the model. However, for clarity of exposition, we will present the earliest 'bipartite' version of GSC, which deals only with disorders and symptoms.

[5] If there is in fact a single disorder which can explain all symptoms, the singletone set containing that disorder as its sole member will, of course, be a candidate.

[6] This set should not be confused with the function *manifs* described above.

[7] Where the disorders are abbreviated to their initials. 'bac' = bad auto choke, etc.

[8] Which has a cumbersome definition but greatly resembles set intersection followed by a Cartesian product of the 'remainders' not included in the intersection.

[9] *Parsimony* is defined as 'minimum cardinality' in the early GSC work and has been more recently amended to 'irredundancy'. This will be discussed in Section 3.

[10] *INTERNIST-I*, whose Bayesian basis is essentially vindicated by Charniak's analysis, did not use intermediate pathological states and was therefore statistically unsound in that respect.

[11] Whether or not they do so, and how, is a separate issue.

[12] Intermediate pathological states or parameters still unexplained or only partially accounted for.

[13] Hierarchies of disorders, based on various views: organ systems involved, mechanisms of operation inside the body, etc. This approach is also being followed in recent work by Kautz (Kautz, 1986).

[14] A key point is that the causal strength is *not* the same as $P(m_i \mid d)$, the posterior probability of the manifestation given the presence of the disorder.

References

Allen, J. (1982) Recognizing intentions from natural language utterances. In *Computational Models of Discourse* (ed. M. Brady), MIT Press, Cambridge MA.

Ben-Bassat, M. Carlson, R.W. Puri, *et. al.* (1980) Pattern based interactive diagnosis of multiple disorders: The *MEDAS* system. *IEEE Transations on Pattern Analysis and Machine Intelligence*, vol. 148–160.

Buchanan, B.G. & Shortliffe, E.H. (ed)(1984) *Rule-Based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA.

Carberry, S. (1987) Plan recognition and its use in understanding dialogue. In *User Models in Dialog Systems* (eds. A. Kobsa & W. Wahlster) Springer-Verlag, Berlin.

Charniak, E. (1983) The Bayesian basis of common sense medical diagnosis. In *Proceedings of the Third National Conference on Artificial Intelligence* pp. 70–73. William Kaufman, Inc., Los Altos, California.

deKleer, J. & Williams, B.C. (1986a) *Diagnosing Multiple Faults.* Technical Report, Xerox PARC, Palo Alto, Calif.

deKleer, J. & Williams, B.C. (1986b) Reasoning about multiple faults. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 132–139. Morgan Kaufmann, Palo Alto, Calif.

Finin, T. (1983) Help and advice in task oriented systems. In *Eighth International Conference on Artifical Intelligence.* International Joint Conference on Artificial Intelligence. Morgan Kaufmann, Calif.

Finin, T., Joshi, A. & Webber, B. (1986) Natural language interactions with artificial experts. *Proceedings of the IEEE* **74**(7), pp. 921–938.

Genesereth, M. (1979) The role of plans in automated consultation. In *Sixth International Conference on Artifical Intelligence*, pp.311–319, Morgan Kaufmann, Palo Alto, Calif.

Genesereth, M.R. (1984) The use of design descriptions in automated idagnosis. *Artifical Intelligence Journal*, **24**, pp. 411–436.

Kaplan, J. (1982) Cooperative responses from a portable natural language database query system. In *Computational Models of Discourse* (ed. M. Brady) MIT Press, Cambridge.

Kassirer, J.P. & Gorry, G.A. (1978) Clinical problem solving: a behavioural analysis. *Annals of Internal Medicine*, **89**.

Kautz, H. (1985) *Toward a theory of plan recognition*. Technical Report TR162, Department of Computer Science, University of Rochester, Rochester, NY.

Kautz, H. (1986) Generalized plan recognition. In *Proceedings of the Fifth National Conference on Artificial Intelligence* pp. 32–37. Morgan Kaufmann, Palo Alto, Calif.

Patil, R.S. (1981) *Causal Representation of Patient Illness for Electrolyte and Acid-Base Diagnosis* Technical Report MIT/LCS/TR-267, Massachussetts Institute of Technology, Laboratory for Computer Science.

Peng, Y. & Reggia, J.A. (1986) Plausibility of diagnostic hypotheses: the nature of simplicity. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 140–145. Morgan Kaufmann, Palo Alto, Calif.

Peng Y. & Reggia J.A. (1986a) *Being Comfortable with Plausible Diagnostic Hypotheses*. Technical Report TR-1753, University of Maryland, Computer Science Department.

Peng, Y. & Reggia, J.A. (1987b) A probabilistic causal model for diagnostic problem solving — Part I: integrating symbolic causal inference with numeric probabilistic inference. *IEEE Transactions on Systems, Man & Cybernetics*, **17**.

Pollack, M.E. (1986) *Inferring Domain Plans in Question-Answering*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

Pollack, M., Hirschber, J. & Webber, B. (1982) User participation in the reasoning processes of expert systems. In *Proceedings of the Second National Conference on Artificial Intelligence*. CMU, Pittsburgh PA. A longer version appears as Technical Report CIS-82-9, Department of Computer and Information Science, University of Pennsylvania, July 1982.

Pople, Jr., H.E. (1973) On the mechanization of abductive logic. In *Third International Conference on Artificial Intelligence*. pp. 147–152. Morgan Kaufmann, Palo Alto, Calif.

Pople, Jr., H.E. (1977) The formation of composite hypotheses in diagnostic problem-solving: an exercise in synthetic reasoning. In *Fifth International Conference on Artifical Intelligence*, pp. 1030–1039. Morgan Kaufmann, Alto, Calif.

Pople, Jr., H.E. (1982) Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnoses. In *AI in Medicine*, (ed. P. Szolovits) pp. 119–190. Westview Press.

Ramsay, C.L., Reggia, J.A., Nau, D.S. & Ferrentino, A. (1986) A comparative analysis of methods for expert systems. *International Journal of Man-Machine Studies*, **24**, 475–499.

Reggia, J.A. (1985) Abductive inference. In *Proceedings of the Expert Systems in Government Symposium*, pp. 484–489.

Reggia, J.A. & Peng, Y. (1986) Modeling diagnostic reasoning: a summary of parsimonious covering theory. In *Proceedings of the IEEE Conference on Computer Applications in Medical Care*, pp. 17–29.

Reggia, J.A., Nau, D.S., & Wang, P. (1985a) A formal model of diagnostic inference — part 1: problem formulation and decomposition. *Information Sciences*, **37**, pp. 227–256.

Reggia, J.A., Nau, D.S. & Want, P. (1985b) A formal model of diagnostic inference — part 2: algorithmic solution and application. *Information Sciences*, **37**, 257–285.

Reiter, R. (1985) *A Theory of Diagnosis from First Principles*. Technical Report TR-187/86, Department of Computer Science, University of Toronto.

Schank, R. & Abelson, R. (1977) *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, Hillsdale, NJ.

Shortliffe, E.H. (1976) *Computer-Based Medical Consultations: MYCIN*. Elsevier, New York.

Shrager, J. & Finin, T. (1982) An expert system that volunteers advice. In *Proceedings of the Second National Conference on Artificial Intelligence*, pp. 339–340. Morgan Kaufmann, Palo Alto, Calif.

Swartout, W. (1983) *XPLAIN*: A system for creating and explaining expert consulting programs. *Artifical Intelligence*, **21**, pp. 285–325.

Weiss, S.M. & Kulikowski, C.A. (1984) *A Practical Guide to Designing Expert Systems*. Rowman and Allanheld, Publishers.

Wilenski, R. (1983) *Planning and Understanding*. Addison-Wesley, Reading, Mass.