# Explaining by Evidence

Eliezer L. Lozinskii
Institute of Computer Science
The Hebrew University, Jerusalem 91904, Israel
email: lozinski@cs.huji.ac.il, tel: 972-2-6585490, fax: 972-2-6585439

**Abstract.** Given a system description and its observed behaviour, there usually exist several possible explanations of the situation. Different explanations may be contradicting and suggesting conflicting actions. So, in order to take proper actions sanctioned by the real situation, we need to single out an explanation that is most adequate to the reality. Existing methods for determining of such a *best explanation* require an extensive statistical knowledge about the system behaviour which often is either only partially available or not satisfactorily reliable. Trying to overcome this lack of data, we introduce the notion of *evidence* as a measure of credibility of beliefs based on the semantic information contained in the system and supplied by the observations. Reasoning by evidence is applied to the process of determining a best abductive explanation. The notion of conjecture is presented as a generalization of explanation. Then reasoning by evidence is applied to testing and refining of diagnoses.

**Keywords:** Abduction, explanation, diagnosis, conjecture, testing, reasoning by evidence.

## 1 Introduction

Presently an extensive research is being devoted to investigating *abduction*, that is, a process of finding reasons, causes and explanations of various events. Introduced by Peirce [5], abduction underlies numerous processes of a great practical importance, such as medical diagnosing [2, 20, 57, 62, 68, 70, 83], testing and repairing technical devices and structures [4, 11, 13, 21, 24, 40, 55, 58, 72], planning a course of action [7, 18, 59], natural language understanding [32, 33, 43, 44, 39, 74, 75], learning [14, 61, 74, 75].

Let $S$ denote a Knowledge System (presented in the language of First Order Logic) that describes a part of a real world $W$. Suppose that a state $Obs$ has been observed in $W$, and we wish to figure out what event or state of affairs caused $Obs$ or accounts for its occurrence. To do so we have to resort to the knowledge about $W$ presented in $S$ in order to find an *explanation* of $Obs$. Several questions arise with regard to the notion of explanation. What accounts for an explanation? What determines the quality of an explanation? What is a best explanation? What is a process of inference to the best explanation? [48]

Numerous works on abduction adopt the following general definition of explanation [18, 19, 46, 58, 59, 66, 76].

**Definition 1.1** *Given a system $S$, and an observation Obs, an* explanation *of Obs in $S$ is a formula $expl(S, Obs)$ that satisfies the following conditions:*
*(i) $S \cup \{expl(S, Obs)\} \models Obs$;*
*(ii) $expl(S, Obs)$ is consistent with $S$.*

Usually several refinements of this definition are considered.

(1) For practical reasons not all formulas of $S$ are most suitable for expressing explanations, but only those (called *abducible formulas*) that describe possible actual causes of observed events or elements of the system amenable to a purposeful treatment, such as known diseases in case of medical diagnosis, or repairable elements of technical devices. For example, if a computer fails in a classroom, then it would be appropriate to detect a faulty board in order to replace it by a good one. In this case abducible formulas are those describing functioning of the computer boards. But when a faulty board is brought to a laboratory for repair, then a faulty chip is to be detected, so, abducible formulas should describe the inner structure of the board. At the production level abducible formulas may refer to individual gates of chips, etc. So, any explanation should be expressed in terms of abducible formulas. Let $AF$ denote a set of all abducible formulas.

(2) An explanation should not be redundant, and has to possess a certain property of parsimony or simplicity [3, 41, 65]. In particular, if there is a set of possible explanations, then each individual explanation is preferable to a disjunction of several ones. For instance, if $S = \{p_1 \to r,\ p_2 \to r\}$, $AF = \{p_1,\ p_2,\ p_3\}$, $Obs = \{r\}$, then $p_1$ (as well as $p_2$ ) is a non-redundant explanation of $r$ , while $p_1 \vee p_2$ is redundant. On the other hand, an explanation should not be *overspecified* [6, 79] in the sense that an explanation has to be preferred to its conjunction with any other statement. In the previous example $p_1$ is an explanation of $r$ preferred to conjunctions $p_1 \wedge p_2$, $p_1 \wedge p_3$ that are overspecified versions of $p_1$, and so, redundant. This intuition suggests the following definition of a desirable property of explanations.

**Definition 1.2** *Let $\sigma$ be a subset of abducible formulas $AF$ involved in an explanation. Then the explanation expressed by the formulas of $\sigma$ is* non-redundant *if there exists no other explanation of the same observation expressed by $\sigma' \subseteq AF$ such that $\sigma' \subset \sigma$ ($\subset$ denotes a proper subset).*

(3) On the other hand, an explanation should not be *vacuous* in the sense that $S \models Obs$, that is, the observation follows directly from the system description, and so, needs no additional explanation. Indeed, in this case for any formula $\phi$, including an empty one, $S \cup \{\phi\} \models Obs$ as $\models$ is a monotonous relation.

(4) Explanations should be *non-trivial*, which means that the observation or its part should not be considered as its own explanation without any use of the system description. For example, if $Obs = \{p \vee q\}$, then neither $p$ nor $q$ is an explanation.

The following definition extends Definition 1.1 to include refinements (1)–(4).

**Definition 1.3 (Strict explanation)** *Given a system description $S$, a set of abducible formulas $AF$, and an observation Obs, a* strict explanation *of Obs in $S$, $expl(S, AF, Obs)$, satisfies the following conditions:*

*(i)* $expl(S, AF, Obs)$ *is expressed in terms of* $AF$*;*
*(ii)* $S \cup \{expl(S, AF, Obs)\} \models Obs$*;*
*(iii)* $expl(S, AF, Obs)$ *is consistent with* $S$*;*
*(iv)* $expl(S, AF, Obs)$ *is not vacuous;*
*(v)* $expl(S, AF, Obs)$ *is non-trivial;*
*(vi)* $expl(S, AF, Obs)$ *is non-redundant.*

It has to be admitted that Definitions 1.1, 1.3 may become a subject of argument, as there are researchers who agree to accept as explanations statements that do not necessarily imply the observation. According to Gärdenfors [23], *expl* is an explanation of *Obs* due to a given state of knowledge $S$ if adopting *expl* would increase the probability of *Obs*, that is, $prob(Obs \mid S \cup \{expl\}) > prob(Obs \mid S)$. Each explanation possesses an *explanatory power*, $EP(expl, Obs)$, that is defined in [23] as a difference between the posterior and prior probabilities of the explanandum, $prob(Obs \mid S \cup \{expl\}) - prob(Obs \mid S)$, and in [6] as their ratio, $prob(Obs \mid S \cup \{expl\}) \ / \ prob(Obs \mid S)$. If there exist several explanations of the same observation, then due to [23] the one with the maximum explanatory power should be preferred to the others. So, this approach would favour an explanation $\alpha$ implying the observation, since $prob(Obs \mid S \cup \{\alpha\})$ reaches its maximum value of 1.

Bearing all that in mind, we shall consider in the sequel strict explanations unless stated otherwise.

Usually there exists a set $EXPL(S, AF, Obs)$ of different explanations of *Obs* in *S*, such that each particular explanation may have its specific consequences, and sanction a certain action often incompatible with the actions suggested by other members of the set. So, to take a proper action, it is important to determine the actual explanation of the observed situation, or at least, be able to prefer an explanation which is most likely to be the actual one. Following sections present an approach to selecting a best explanation based on the notion of semantic information, entropy and evidence.

# 2 Best explanation

Selecting a best explanation depends on several criteria and circumstances. We may either prefer a most likely explanation conforming to the common sense intuition, or select the one requiring a most urgent action, or follow up an explanation suggesting an easiest or cheapest treatment, etc. Poole and Provan [67] have studied widely adopted approaches to diagnostic reasoning, and compared the associated notions of optimality. They have shown that the existing approaches select different best explanations which often are incompatible.

Preferring an explanation with a highest explanatory power [23] is one way of determining a best explanation. However, as has been shown in [6], this approach may miss the real cause of a given observation since it does not take into account the prior probabilities of conflicting explanations of the observation. As an alternative way of choosing a best explanation, Chajewska and Halpern [6] consider both the explanatory power of an explanation, $EP(expl,\ Obs)$ and its prior probability, $prob(expl)$. They define a partial order on explanations of $Obs$ such that an explanation $\alpha$ is preferred to $\beta$ if $EP(\alpha,\ Obs) > EP(\beta,\ Obs)$ and $prob(\alpha) \geq prob(\beta)$ or $EP(\alpha,\ Obs) \geq EP(\beta,\ Obs)$ and $prob(\alpha) > prob(\beta)$. Unfortunately, this approach leaves many pairs of competing explanations incomparable.

An approach adopted by many researchers [8, 9, 31, 63, 64, 66, 79, 82], called *Most Probable Explanation (MPE)* in [63] or *Maximum A-Posteriory (MAP)* assignment in [8, 9, 79], is selecting an explanation *expl* possessing a highest probability given all the available knowledge, that is, maximizing the posterior probability $prob(expl|S \cup Obs)$. This approach is based on the *Bayes' Theorem* that allows computing unknown probability of some event given corresponding probabilities of other ones [45, 63]. Let $p(A)$, $p(A|B)$ denote, respectively, the prior probability that a statement (formula) $A$ is true, and the posterior probability that $A$ is true given that $B$ is true. If $p(A) \neq 0$ and $B_1, B_2, \ldots, B_n$ are mutually exclusive and jointly exhaustive statements such that $\sum_{i=1}^{n} p(B_i) = 1$, then

$$p(B_i|A) = \frac{p(B_i)p(A|B_i)}{p(A)} = \frac{p(B_i)p(A|B_i)}{\sum_{i=1}^{n} p(B_i)p(A|B_i)}.$$

To apply the Bayesian approach to computing of $p(B_i|A)$ one needs to know $p(B_i)$ and $p(A|B_i)$ for all $i = 1, 2, \ldots, n$. This requirement of a full specification of probability distribution is often difficult to meet in real systems.

## 2.1 Evidence *versus* probability

Several methods have been developed to cope with a lack of statistical information. A celebrated one is the *Principle of Maximum Entropy* [35, 36, 84]. The notion of entropy, widely used in physics, was considered by Hartley [30] and Wiener [86] in connection with information transmission, and introduced into Information Theory by Claude Shannon [78]. Let $P$ be a probability distribution assigning probabilities $p_1, \ldots, p_n$ to $n$ mutually exclusive and jointly exhaustive events (or sentences) of a set $R$. Then the *entropy* $H(P)$ of $P$ is defined by

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i, \tag{1}$$

where $p_i \log p_i$ is assigned zero if $p_i = 0$.

$H(P)$ is a measure of uncertainty associated with $R$ given the distribution $P$. In particular, $H(P)$ reaches its maximum value of $\log n$ if all the events of $R$ are equiprobable,

$$p_1 = \ldots = p_n = 1/n. \tag{2}$$

This distribution is, indeed, most uncertain, since it does not allow to expect one event more than any other, or to prefer the truth of some statement of $R$ to that of another one.

Now suppose that some knowledge $K$ is available about real relationships existing among the elements of $R$. This knowledge can be regarded as a set of constraints on $R$. As the available knowledge about the real world is usually incomplete, there exists a set $\{P_K\}$ of different distributions satisfying the constraints. Then an important question is, which one of the various $P_K$ is most adequate to the reality? This *preferred distribution*, $P_K^*$, must reflect the fact that there is no other knowledge about $R$ beyond that presented by $K$. So, besides satisfying the constraints, $P_K^*$ must maximize the uncertainty associated with $R \cup K$. And this is just the decision suggested by the Principle of Maximum Entropy: Given $R$ and $K$, prefer a probability distribution $P_K^*$ satisfying $K$ such that for all $P_K$, $H(P_K^*) \geq H(P_K)$. This principle, which can be traced back some 100 years, to the works of Boltzmann, Maxwell, Gibbs, and

Planck, and has been introduced in its modern form by Jaynes [35] and Tribus [84], proved its efficiency in various fields of Physics, Statistics, Information Theory, Pattern Recognition, Signal Processing, etc. [1, 28, 29, 35, 36, 37, 38, 81, 84, 87].

To be applied successfully, most of the known approaches to reasoning with uncertain and incomplete information require an extensive statistical knowledge (e.g., prior and posterior probabilities for Bayesian methods, basic probability assignment in Dempster-Shafer theory [15, 16, 77], prior beliefs for computing certainty factors in MYCIN [31, 80], partial probability assignments for Nilsson's Probabilistic Logic [60]). In many practical situations such statistical data are either only partially available or unsatisfactorily reliable. These circumstances raise an important question of how reasoning with incomplete information can be performed efficiently in the absence of necessary statistical data.

Although the available knowledge $S$ may not imply the truth or falsehood of a statement $F$, it provides some *semantic information* about $F$. For instance, if $S = \{p \vee q\}$, then $S$ suggests that $p$ and $q$ cannot be false together, so, if $q$ is false, then $p$ is certainly true, etc. As $S$ enables reaching certain conclusions regarding the truth of $F$, we may say that $S$ provides certain *evidence* of $F$ (denoted $E(S, F)$). We would like the value of evidence $E(S, F)$ to present a measure of credibility of a belief in the truth of $F$. Actually, we need to know the probability that $F$ is true in a given state of the world described by $S$, but if it is impossible to compute the probability of $F$, in particular, because of a lack of necessary statistical data, then we would resort to the value of $E(S, F)$. Evidence, considered as a surrogate for probability, is different from the latter both conceptually and quantitatively. But since it is intended to substitute probability in the process of reasoning, evidence has to possess the following properties conforming to the axioms of probability.

(1) If $F$ is a logical consequence of $S$, then the truth of $F$ is most credible, so $E(S, F)|_{S \models F} = 1$. (Let the range of evidence be normalized in [0, 1]).

(2) If $S \models \neg F$, then the belief in the truth of $F$ has no credibility. Hence, $E(S, F)|_{S \models \neg F} = 0$.

(3) In particular, if $\alpha, \beta$ are a tautology and an unsatisfiable formula, respectively, then for all $S$, $E(S, \alpha) = 1$ and $E(S, \beta) = 0$.

(4) For all formulas $\phi, \psi$, $E(S, \phi \vee \psi) = E(S, \phi) + E(S, \psi) - E(S, \phi \wedge \psi)$.

(5) Properties (3) and (4) imply for all $S$ and $\psi$, $E(S, \psi) + E(S, \neg\psi) = 1$.

If $S$ represents a world $W$ faithfully, then every momentary state of $W$ corresponds to a model of $S$, although, because of an incompleteness of the knowledge presented in $S$, some models may describe states of $W$ that do not occur in the reality. However, these states appear to be possible from the standpoint of the current knowledge. This observation underlies the widely adopted view of $W$ as a set of states, referred to as *possible worlds*, such that there is a bijection between the set of possible worlds and the set $MOD(S)$ of all models of $S$ ($S$ is supposed to be consistent).

At any moment the world is in one and only one of its possible states. So, $MOD(S)$ provides a set of mutually exclusive and exhaustive particular representations of $W$. Let $p(\mu)$ denote the probability that $W$ is in a state represented by a model $\mu \in MOD(S)$. Then the probability $p(F)$ that a formula $F$ is true in the current state of $W$ can be expressed as

$$p(F) = \sum_{\mu \in MOD(S)} p(F|\mu)p(\mu) = \sum_{\mu \in MOD(S) \wedge \mu \models F} p(\mu), \tag{3}$$

since

$$p(F|\mu) = \begin{cases} 1 & if \;\; \mu \models F \\ 0 & otherwise. \end{cases}$$

Unfortunately, a full probability distribution over the space of possible states is usually either not available or not satisfactorily reliable, although a partial distribution may be known for some possible worlds represented by a subset of models $M \subset MOD(S)$. If $M$ does not contain all models asserting $F$, then a question arises of how the probability of the models not contained in $M$ should be estimated.

## 2.2 Applying the Principle of Maximum Entropy

Consider a user interested in finding out the actual state of a world $W$ given its description by a knowledge system $S$. As the system represents every possible world by means of one of its models, a probability distribution over $MOD(S)$ is necessary for reasoning about the actual state of $W$. Suppose that prior probabilities $\rho(\mu)$ are known for models of a subset $M$ of $MOD(S)$. Given this partial distribution $P_M$, a rational reasoner should prefer a full distribution $P_{MOD}$ over $MOD(S)$ that (a) is consistent with $S$, (b) includes $P_M$, and (c) implies no extra knowledge beyond $S$ and $P_M$. According to the Principle of Maximum Entropy [35, 84], $P_{MOD}$ must maximize the value of Shannon's entropy $H(P)$ (expression (1)). This implies the following definition of $P_{MOD}$:

$$p(\mu) = \begin{cases} \rho(\mu) & if \; \mu \in M \\ \frac{1-\sum_{\mu \in M} \rho(\mu)}{|MOD(S)-M|} & otherwise. \end{cases} \tag{4}$$

Now the evidence $E(S, F)$ can be defined as the probability that $F$ is true in the current state of the world given the reasonable, although not necessarily actual, distribution $P_{MOD}$.

**Definition 2.1 (Evidence)** *Due to expressions (3) and (4),*

$$\begin{aligned} E(S,F) &= \sum_{\mu \in M \wedge \mu \models F} \rho(\mu) + \sum_{\mu \in (MOD(S)-M) \wedge \mu \models F} p(\mu) \\ &= \sum_{\mu \in (M \cap MOD(S \cup \{F\}))} \rho(\mu) + (1 - \sum_{\mu \in M} \rho(\mu)) \frac{|MOD(S \cup \{F\}) - M|}{|MOD(S) - M|}, \end{aligned} \tag{5}$$

*where $|MOD(S)|$, $|MOD(S \cup \{F\})|$ denote, respectively, the cardinality of the set $MOD(S)$ of all models of $S$, and that of its subset consisting of all models in which $F$ is true.*

*If $MOD(S \cup \{F\}) \subseteq M$, then $E(S, F)$ amounts to the probability of $F$. If no prior probabilities of possible worlds are known, such that $M = \emptyset$, then*

$$E(S,F) = \frac{|MOD(S \cup \{F\})|}{|MOD(S)|}, \tag{6}$$

*that corresponds to the assumption (by the Principle of Maximum Entropy) that all possible worlds are equiprobable.*

It is easy to see that Definition 2.1 is in accord with properties (1)-(5) of evidence (Subsection 2.1).

An implicit assumption underlying Definition 2.1 is that the set of models of $S$ is finite. However, it has been shown in [52] how the value of evidence can be calculated in systems with infinite sets of models.

## 2.3 Reasoning by evidence

The preceding considerations suggest the following way of reasoning by adopting evidence $E(S, F)$ as a measure of credibility of a belief in the truth of $F$ given $S$.

**Definition 2.2 (Reasoning by evidence)** *(i) If $E(S, F) = 1$ (or $E(S, F) = 0$), then $F$ is true (false, respectively);*

*(ii) If $E(S, F) > 0.5$ (that is, $E(S, F) > E(S, \neg F)$), then $F$ is more likely to be true than false, and the larger the value of $E(S, F)$ the more credible a belief in the truth of $F$; If $E(S, F) = 0.5$, then both possible beliefs of $F$ have the same (rather low) credibility.*

*(iii) If $E(S, F) > E(S, F')$, then the truth of $F$ is more credible than that of $F'$.*

In situations where one has to resort to reasoning by evidence for assessing the relative quality of possible explanations, she or he would favour an explanation possessing a higher evidence.

**Definition 2.3 (Best explanation)** *Given a system description $S$, an observation $Obs$, and a set $EXPL$ of explanations of $Obs$ in $S$,*

*(i) an explanation $\phi \in EXPL$ is* better *than $\psi \in EXPL$ if $E(S \cup Obs, \phi) > E(S \cup Obs, \psi)$;*

*(ii) a* best explanation *is $\alpha \in EXPL$ such that*

$$E(S \cup Obs, \alpha) = \max_{expl \in EXPL} [E(S \cup Obs, expl)].$$

It turns out that if an explanation $\phi$ of $Obs$ is better (by Definition 2.3(i)) than another explanation $\psi$ of $Obs$ , then in $S$ itself $\phi$ possessed an evidence greater than that of $\psi$ even before the observation has been made.

**Proposition 2.1** *For all $\phi, \psi \in EXPL$,*

$$E(S \cup Obs, \phi) \geq E(S \cup Obs, \psi) \quad iff \quad E(S, \phi) \geq E(S, \psi).$$

**Proof.** By Definition 1.3(ii),

$$MOD(S \cup \{\phi\}) \subseteq MOD(Obs); \quad MOD(S \cup \{\psi\}) \subseteq MOD(Obs).$$

Hence,

$$MOD(S \cup \{\phi\}) = MOD(S \cup \{\phi\} \cup Obs); \quad MOD(S \cup \{\psi\}) = MOD(S \cup \{\psi\} \cup Obs).$$

If $E(S \cup Obs, \phi) \geq E(S \cup Obs, \psi)$, then $|MOD(S \cup Obs \cup \{\phi\})| \geq |MOD(S \cup Obs \cup \{\psi\})|$, so, $|MOD(S \cup \{\phi\})| \geq |MOD(S \cup \{\psi\})|$, and hence, $E(S, \phi) \geq E(S, \psi)$.

By the same argument, if $E(S \cup Obs, \phi) < E(S \cup Obs, \psi)$, then $E(S, \phi) < E(S, \psi)$.

# 3 An example

Let us apply reasoning by evidence to diagnosing of a binary full adder that is commonly used as a sample target of various abduction methods [24, 58, 72].

Diagnosing a technical device becomes necessary when the device exhibits an abnormal behaviour. In case of an electronic circuit it happens when given a certain inputs, the circuit produces outputs that are different from those corresponding to its normal operation. To repair the device one has to detect its faulty components that caused the abnormal behaviour. And this is the aim of diagnosing. So, a diagnosis is an explanation of the observed behaviour expressed by means of a conjunction $\Delta$ of faulty components, while the rest of the device is assumed to operate correctly. A statement that all system components are operating normally *by default* (unless being involved in $\Delta$ ) is usually included in the system description. This makes $\Delta$ conform to Definition 1.3 of explanation.

Figure 1 displays a diagram of a binary full adder with three binary inputs $x, y, z$, and two binary outputs: $s$ (*sum*) and $c$ (*carry*). A normal operation of the adder is described by table 1. The adder is composed of five elementary *gates* performing Boolean functions on their inputs. $A1, A2$ denote $AND$-gates (performing conjunction of the inputs), $O1$ is an $OR$-gate (disjunction), $X1, X2$ are $XOR$-gates (exclusive-OR). Table 2 shows a normal operation of these gates.

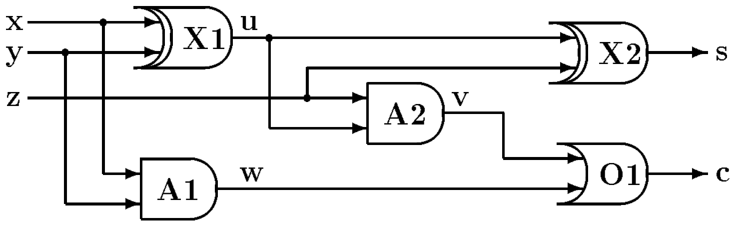


**Figure 1.** Binary full adder.

**Table 1**

| x | y | z | s | c |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 0 | 1 |
| 1 | 1 | 1 | 1 | 1 |

**Table 2**

| Input 1 | Input 2 | Output | | |
|---|---|---|---|---|
| | | AND | OR | XOR |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 0 |

A description of the adder consists of the contents of tables 1 and 2 together with a description of connections between the gates shown in figure 1. If a gate fails to operate normally, it may cause an abnormal operation of the adder. A typical failure state of a gate is such that the gate's output remains permanently 0 or 1 regardless of the values of the gate's inputs. In this case an abnormal gate $G$ is said to be *stuck-at-0* (denoted $s0(G)$) or *stuck-at-1* ($s1(G)$). So, abducible formulas related to the adder of figure 1 are

$$AF = \{s0(G), s1(G) | G \in \{A1, A2, O1, X1, X2\}\}. \tag{7}$$

Suppose that the adder's input is $x = y = z = 0$ (which becomes a part of the system description), while the observed output is $Obs = \{s = 1 \wedge c = 1\}$, which is abnormal as the correct one should be $s = c = 0$. Given these data, there are 6 minimal cardinality diagnoses (explanations), each specifying two faulty gates:

$$\begin{array}{ll} EXPL = \{ & \Delta_1 = s1(X2) \wedge s1(O1),\ \Delta_2 = s1(X2) \wedge s1(A1), \\ & \Delta_3 = s1(X2) \wedge s1(A2),\ \Delta_4 = s1(X1) \wedge s1(O1), \\ & \Delta_5 = s1(X1) \wedge s1(A1),\ \Delta_6 = s1(X1) \wedge s1(A2)\ \ \}. \end{array} \tag{8}$$

Given a diagnosis, certain actions have to be performed in order to restore a normal operation, such as repairing or replacing the faulty components of a device, or administering certain medicine to a patient. However, different diagnoses may sanction different or even contradicting actions. So, to determine a reasonable action, the set of diagnoses has to be disambiguated, that is, a specific diagnosis should be singled out of the set as the best one.

Possible diagnoses may differ significantly in the seriousness of the corresponding malfunction. This consideration may dominate other parameters of diagnoses. For instance, if a diagnosis indicates a possibility of an extreme damage, then all measures must be taken to prevent the grave consequences even if the diagnosis is rather unlikely. We consider a less critical situation in which all possible diagnoses bear approximately the same urgency (like in the case of a faulty binary adder).

Poole and Provan [67] have described and analysed six common approaches to determining a best diagnosis.

(1) Most likely single-fault hypothesis approach: seeks a diagnosis $H$ with a highest probability $prob(H|S \cup Obs)$ among all diagnoses involving a single faulty component, while the rest of the system is assumed to operate correctly.

(2) Most likely posterior hypothesis: prefers a diagnosis $h$ involving a single faulty component regardless of the state of all other components, such that $prob(h|S \cup Obs)$ has a maximum value among all possible single faults.

(3) Most likely interpretation: finds an interpretation $I$ (that is an assignment of truth values to all abducible formulas) with a highest probability $prob(I|S \cup Obs)$ among all possible interpretations implying $Obs$.

(4) Most likely consistency-based diagnosis: determines a most probable minimal set of faulty components which is consistent with $S \cup Obs$, while all other components operate correctly.

(5) Most likely explanation: prefers an explanation $expl$ with a highest probability $prob(expl|S \cup Obs)$.

(6) Utility-based explanation: seeks a conjunction of hypotheses about faulty components that is most useful from the standpoint of the suggested treatment.

These approaches require extensive statistical data that often are either incomplete or not sufficiently reliable. It turns out that in the absence of a necessary statistics the approach of reasoning by evidence provides means for estimation of a relative likelihood of different explanations.

To illustrate reasoning by evidence, let us now return to the example of a binary adder with $x = y = z = 0$, $s = c = 1$, and try to figure out which of the six diagnoses presented in $EXPL$ is the most credible one. Consider the set $AF$ (expression (7)) of all abducible formulas. As every gate can either operate normally or be in a state of $s0$ or $s1$, there are $3^5 = 243$ interpretations of $AF$, while 56 of them satisfy $S \cup Obs$, so, $|MOD(S \cup Obs)| = 56$. Table 3 presents the cardinality of $MOD(S \cup Obs \cup \Delta_i)$, and the value of $E(S \cup Obs, \Delta_i)$, for $i = 1, \ldots, 6$.

**Table 3**

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\lvert MOD(S \cup Obs \cup \Delta_i)\rvert$ | 27 | 18 | 18 | 18 | 12 | 12 |
| $E(S \cup Obs, \Delta_i)$ | 0.482 | 0.321 | 0.321 | 0.321 | 0.214 | 0.214 |

So, due to reasoning by evidence (Definition 2.2), $\Delta_1 = s1(X2) \wedge s1(O1)$ is the most likely diagnosis. This choice appeals to a common sense intuition. Indeed, if $X2$ and $O1$ are stuck-to-1, then $s = c = 1$ regardless of the state of $X1, A1, A2$, while other diagnoses restrict possible states of the components. For instance, if we adopt $\Delta_2 = s1(X2) \wedge s1(A1)$ then given the observed output, $O1$ cannot be stuck-to-0.

The values of evidence in table 3 were calculated without taking into account any statistics regarding the components' behaviour. To compare the diagnoses probabilistically, let us assume that both possible failures of each gate occur with the same probability $p$ independently of the state of the rest of the circuit. Then every model $\mu \in MOD(S \cup Obs)$ with $f(\mu)$ faulty components has a posterior probability

$$prob(\mu|S \cup Obs) = \frac{p^{f(\mu)}(1-2p)^{5-f(\mu)}}{\sum\limits_{m \in MOD(S \cup Obs)} p^{f(m)}(1-2p)^{5-f(m)}} \tag{9}$$

All models are mutually exclusive, so, given $Obs$, a posterior probability of $\Delta_i$ is

$$prob(\Delta_i|S \cup Obs) = \sum_{\mu \in MOD(S \cup Obs \cup \{\Delta_i\})} prob(\mu|S \cup Obs) \tag{10}$$

Table 4 presents probabilities of the six diagnoses approximated to the first power of $p$. It shows that $\Delta_1$ is the most probable diagnosis, while $\Delta_5$ and $\Delta_6$ are the least probable ones, which is in accord with the estimation made by means of reasoning by evidence (table 3).

**Table 4**

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $prob(\Delta_i \vert S \cup Obs)$ | $.17 + .44p$ | $.17 + .28p$ | $.17 + .28p$ | $.17 + .28p$ | $.17 + .11p$ | $.17 + .11p$ |

# 4 Conjecture *versus* explanation

Producing a strict explanation of a given observation requires a rather extensive knowledge about the system and its abducible formulas. If the available knowledge is insufficient, it may happen that there is no formula that satisfies all the requirements of Definition 1.3. Nevertheless, the available knowledge may suggest some conjecture regarding the observed situation. An important question is *"How can such a reasonable conjecture be produced out of an incomplete knowledge?"* This section will address this issue.

Recently several research works [7, 18, 59] have applied abduction to the process of planning. Indeed, suppose that we wish to achieve a goal $G$ in a system $S$. We may consider $G$ as a desirable observation. Let $Op$ denote a set of possible operations, that should in the framework of abduction be represented by a set of abducible formulas of $S$. Then (by Definition 1.3) $expl(S, Op, G)$ determines actions that are necessary and sufficient for achieving the goal.

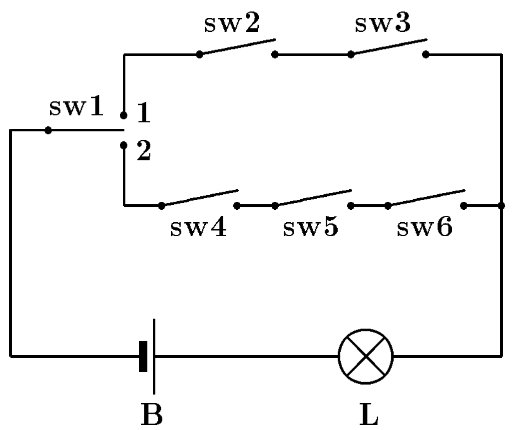


**Figure 2.** Electric circuit.

Let us consider a simplistic motivating example of an electric circuit displayed in figure 2. It is composed of a light bulb $L$, battery $B$, and switches $sw1, \ldots, sw6$. Normally the battery and the bulb are operational, but all the switches are in state *off*, so, the bulb is not lit. There are two ways to lit $L$: either put $sw1$ in state *1* and $sw2, sw3$ in state *on*, or put $sw1$ in state *2* and $sw4, sw5, sw6$ in state *on*. Suppose that $sw1, \ldots, sw6$ are located in different places, but all the switches are equally accessible to an operator, so, in terms of abduction $sw1, \ldots, sw6$ are presented in the set of abducible formulas of the system. To lit the bulb (that is the desirable observation) the operator would prefer to operate three switches $sw1, sw2, sw3$ rather than four ones $sw1, sw4, sw5, sw6$. Now suppose that $sw2, sw3$ and $sw6$ become (temporarily) inaccessible to the operator, such that only $sw1, sw4, sw5$ are abducible. It appears reasonable that in this case the operator would put $sw1$ in state *2* and switch on $sw4, sw5$ in hope that it is more likely to get an access to a single switch $sw6$ than to two inaccessible ones $sw2, sw3$. Besides, it is possible that some of the inaccessible switches are already switched on, and again, it is more likely that a single $sw6$ is on than the two $sw2$ and $sw3$ are so.

In general, given a goal $G$, and being unable to achieve it in the present situation, a rational reasoner would perform an action which gives the best chance that the goal will be achieved if and when the circumstances change. The approach of reasoning by evidence suggests that this kind of a chance is characterized by the evidence of $G$ provided by the state of the system resulted from the action. So, the most appropriate is an action providing a maximum evidence of the goal. With the aforementioned correspondence between an observation and a goal, and between an explanation of the observation and an action achieving the goal, it appears that a preparatory action providing a high chance of achieving the goal corresponds to a credible conjecture regarding reasons of the observation.

**Definition 4.1 (Conjecture)** *Given a system description $S$, a set of abducible formulas $AF$, and an observation $Obs$, a* conjecture *regarding reasons of $Obs$ in $S$, $con(S, AF, Obs)$, satisfies the following conditions:*

($i$) *$con(S, AF, Obs)$ is expressed in terms of $AF$;*

($ii$) $E(S \cup \{con(S, AF, Obs)\}, Obs) = \max_{\phi} [E(S \cup \{\phi\}, Obs)]$ *for all formulas $\phi$ expressed in terms of $AF$;*

($iii$) *$con(S, AF, Obs)$ is consistent with $S$;*

($iv$) *$con(S, AF, Obs)$ is not vacuous, that is, $E(S \cup \{con(S, AF, Obs)\}, Obs) > E(S, Obs)$;*

($v$) *$con(S, AF, Obs)$ is non-trivial, that is, $E(S \cup \{con(S, AF, Obs)\}, Obs) > E(con(S, AF, Obs), Obs)$;*

($vi$) *$con(S, AF, Obs)$ is non-redundant in the sense that if it is expressed by a set $\sigma$ of abducible formulas , then there exists no formula $\phi$ expressed by a proper subset $\sigma'$ of $\sigma$ such that $E(S \cup \{\phi\}, Obs) \geq E(S \cup \{con(S, AF, Obs)\}, Obs)$.*

Definition 4.1 presents a generalization of Definition 1.3 such that explanation is an extreme case of the notion of conjecture that provides the observation with a maximum evidence of 1. At the same time (as evidence is used as a surrogate for probability), the evidence of the observation provided by a conjecture, $E(S \cup \{con(S, AF, Obs)\}, Obs)$, is akin to the posterior probability of the observation considered by Gärdenfors [23] as a measure of quality of the corresponding explanation.

Consider again the circuit of figure 2, and let $S$ include a description of the circuit , and an assumption that the battery and the bulb are normal. We wish to light the bulb, that is, make $Obs = On(L)$. If $sw2, sw3, sw6$ are inaccessible, then $AF = \{$*On1(sw1), On2(sw1), Off(sw1), On(sw4), Off(sw4), On(sw5), Off(sw5)*$\}$, where $On1(sw1)$ means that $sw1$ is switched on to state 1. To determine an action providing the best way to lighting the bulb, let us compute the conjecture regarding reasons of $On(L)$ in $S$ given the inaccessibility of $sw2, sw3, sw6$. Table 5 gives values of evidence $E(S \cup \{\phi\}, On(L))$ for all possible combinations $\phi$ of states of the abducible switches (if *sw1* is *off* then obviously the bulb cannot be lit, so, $E(S \cup \{\phi\}, Obs) = 0$

for all $\phi$ containing $Off(sw1)$).

**Table 5**

| $\phi$ | | | $\lvert MOD(S \cup \{\phi\})\rvert$ | $\lvert MOD(S \cup \{\phi\} \cup Obs)\rvert$ | $E(S \cup \{\phi\}, Obs)$ |
|---|---|---|---|---|---|
| sw1 | sw4 | sw5 | | | |
| On1 | On | On | 8 | 2 | 0.25 |
| On1 | On | Off | 8 | 2 | 0.25 |
| On1 | Off | On | 8 | 2 | 0.25 |
| On1 | Off | Off | 8 | 2 | 0.25 |
| On2 | On | On | 8 | 4 | **0.50** |
| On2 | On | Off | 8 | 0 | 0.00 |
| On2 | Off | On | 8 | 0 | 0.00 |
| On2 | Off | Off | 8 | 0 | 0.00 |

It follows that $con(S, AF, Obs) = On2(sw1) \wedge On(sw4) \wedge On(sw5)$. Thus the conjecture suggests the same action as the operator's common sense decision justified above.

# 5 Verifying explanations

Given a system described by $S$ and $AF$, an observation $Obs$ may lead to several different explanations constituting a set $EXPL$. Our goal is finding out the actual cause of the observed system behaviour. So, we would like to determine which one of the possible explanations in $EXPL$ describes most faithfully the cause of $Obs$. Possible explanations usually possess different credibility, and we tend to adopt a most credible one. However, unless the evidence of the most credible explanation is exactly 1.0, another explanation possessing even a relatively low credibility may happen to be the actual one. Yet, preferring a most credible explanation we would feel more confident if its credibility measured by its evidence were sufficiently high absolutely, and not only relatively to other explanations. *So, given a set of explanations, we may take some action to increase the credibility of a preferred one.* Furthermore, different explanations may not necessarily be mutually exclusive. Indeed, different diseases may coexist producing the same symptoms, or different units of a device may fail simultaneously causing the same malfunction. *So, along with trying to increase the credibility of some explanation, we may wish to eliminate or discredit other explanations by decreasing their credibility.*

To achieve these goals we need more information about the present state of the system than that provided by $Obs$. Such an additional information can be acquired by means of making more observations. Consider a formula $\theta$ that describes a certain parameter of the system. If the value of $\theta$ can be measured without changing the present state of the system, call $\theta$ an *observable* formula. Let $OF$ denote a set of all observable formulas of the system.

One way of getting more information about the system in addition to that of $Obs$ is an *active testing*, which consists in changing the state of the system, and checking the new values of observable formulas (usually involved in $Obs$). For example, an electrocardiogram of a patient at rest can be compared to that recorded under a physical effort. Or a malfunctioning unit can be removed from a device, and then checked under different inputs.

Another way of getting additional information is a *passive testing*, that consists in checking observable formulas (not involve in $Obs$) without changing the state of the system. For instance, a blood test can be taken of a resting patient in addition to a cardiogram. Or more parameters of a malfunctioning unit can be measured without disconnecting it from the device.

In many practical cases the changes of a system necessary for an active testing are either not allowed or impossible, while a passive testing of additional parameters can usually be performed harmlessly. In this section we consider passive testing as a means for verifying explanations.

Efficient testing of various systems is the subject of an extensive research [40, 56, 72], and several approaches to generating tests have been developed.

De Kleer and Williams [40] developed methods for diagnosing multiple faults. In particular, they studied troubleshooting of digital circuits, where observations are made by means of measuring various parameters of a circuit. The goal of testing is identifying the actual failure in a minimum number of measurements, while the expected cost and length of this process is determined by the entropy $H = -\sum p_i \log p_i$, where $p_i$ is the probability of a possible diagnosis. Then the best measurement is the one that minimizes the entropy resulting from the measurement. The initial probabilities $p_i$ are computed from empirical data, and then recomputed after each measurement.

Reiter [72] considered measurements as a source of information necessary for discriminating between possible diagnoses, and finding the unique actual one. Proposition 5.5 in [72] proves an important relationship between diagnoses and measurements that can be reformulated in terms of this paper as follows.

**Proposition 5.1** *[72] An explanation $expl(S, AF, Obs)$* predicts *$\theta$ if $S \cup \{expl(S, AF, Obs)\} \models \theta$. Then given $S$ and $AF$, every explanation of $Obs$ that predicts $\theta$ is an explanation of $Obs \cup \{\theta\}$, but not an explanation of $Obs \cup \{\neg\theta\}$.*

So, a measurement returning $\theta =' true'$ retains all explanations of $Obs$ that predict $\theta$, and rejects all explanations predicting $\neg\theta$. Thus, measuring $\theta$ is most efficient if every explanation presented in $EXPL$ predicts either $\theta$ or $\neg\theta$, although this measurement may give rise to a new explanation not contained in $EXPL$. This approach suggests a divide-and-conquer strategy for discriminating among explanations.

McIlraith [56] studied the problem of generating tests. In particular, [56] extended the approach developed by Reiter [72] to active testing. Let a *test* be a pair $< A, \theta >$, where $\theta$ is an observable formula, and $A$ denotes a set of formulas describing special conditions imposed on the system for the purpose of the test (if $A = \emptyset$ then the test is passive). Given an explanation $\alpha = expl(S, AF, Obs)$, and a test $< A, \theta >$ such that $S \cup A \cup \{\alpha\}$ is satisfiable, an outcome $\theta =' true'$ of the test *confirms* $\alpha$ iff $(S \cup A) \models (\alpha \rightarrow \theta)$, but *refutes* $\alpha$ iff $(S \cup A) \models (\alpha \rightarrow \neg\theta)$. Based on this definition, McIlraith developed a method that uses abduction for generating tests intended for discriminating among explanations.

Existing approaches to testing have certain limitations (alas, nothing is perfect). The method based on minimization of entropy (developed in [40]) requires a knowledge about probabilities of possible explanations. Often such an exhaustive statistics is either unavailable or not sufficiently reliable. The testing methods developed in [24, 56, 72] are based on measuring values of the system parameters which are implied by some of the explanations that have to be discriminated. However, in many practical cases (especially with passive testing) the explanations imply no parameter values other than those of the initial observation. For instance, in the case of a malfunctioning adder considered in Section 3, there is no observable formula (of the set $\{s, c, u, v, w\}$), beyond the observed output of the adder, that is implied by the most likely explanation $s1(X2) \wedge s1(O1)$.

It turns out that reasoning by evidence allows to overcome these limitations: it does not require statistical knowledge (but can accommodate available probabilistic data), and exploits any additional information about the system state (not necessarily implied by the explanations) to modify the credibility of possible explanations leading to their discrimination.

An addition of information $\theta$ to a system may change the evidence provided by the system to a formula $\phi$, such that

$$\delta_\phi(\theta) = E(S \cup \{\theta\}, \phi) - E(S, \phi). \tag{11}$$

**Definition 5.1** *Given a system $S$, and formulas $\phi, \psi$ such that $S \cup \{\phi, \psi\}$ is satisfiable, we say that $\psi$* strengthens *$\phi$ if $\delta_\phi(\psi) > 0$, $\psi$* weakens *$\phi$ if $\delta_\phi(\psi) < 0$, $\psi$ is* indifferent *to $\phi$ if $\delta_\phi(\psi) = 0$.*

**Example 5.1** *Given $S = \{a \vee \neg b, \neg a \vee b \vee \neg c\}$, and a formula $\phi = a$, we get $E(S, \phi) = 0.60$. If $S$ is augmented by a formula $\theta$, then the evidence of $\phi$ changes as shown in table 6. So, $b$ strengthens $a$, while $c$ weakens it.*

**Table 6**

| $\theta$ | $b$ | $\neg b$ | $c$ | $\neg c$ |
|---|---|---|---|---|
| $E(S \cup \{\theta\}, \phi)$ | 1.00 | 0.33 | 0.50 | 0.67 |
| $\delta_\phi(\theta)$ | 0.40 | -0.27 | -0.10 | 0.07 |

The following Proposition characterizes mutual influence of formulas.

**Proposition 5.2 (Reciprocity)** *Given a system $S$, and formulas $\phi, \psi$ such that $S \cup \{\phi, \psi\}$ is satisfiable,*

*(a) $\psi$ strengthens $\phi$ (weakens $\phi$ or is indifferent to $\phi$) iff $\phi$ strengthens $\psi$ (weakens $\psi$ or is indifferent to $\psi$, respectively);*

*(b) $\psi$ strengthens $\phi$ (weakens $\phi$ or is indifferent to $\phi$) iff $\neg\psi$ weakens $\phi$ (strengthens $\phi$ or is indifferent to $\phi$, respectively);*

*(c) The larger $|\delta_\phi(\psi)|$, the larger $|\delta_\phi(\neg\psi)|$.*

**Proof.** (a) For all $\diamond \in \{<, =, >\}$, if $\delta_\phi(\psi) \diamond 0$, then (see equation (11))

$$\frac{|MOD(S \cup \{\phi, \psi\})|}{|MOD(S \cup \{\psi\})|} = k \frac{|MOD(S \cup \{\phi\})|}{|MOD(S)|}, \quad \text{where} \quad k \diamond 1.$$

This implies

$$\frac{|MOD(S \cup \{\phi, \psi\})|}{|MOD(S \cup \{\phi\})|} = k \frac{|MOD(S \cup \{\psi\})|}{|MOD(S)|},$$

that is, $\delta_\psi(\phi) \diamond 0$.

(b) By (a),

$$\frac{E(S \cup \{\psi\}, \phi)}{E(S, \phi)} = \frac{E(S \cup \{\phi\}, \psi)}{E(S, \psi)} = k, \tag{12}$$

such that $k > 1$ ($k < 1$ or $k = 1$) in the case that $\phi$ and $\psi$ are mutually strengthening (weakening or indifferent, respectively).

Denote

$$m = \frac{E(S \cup \{\neg\psi\}, \phi)}{E(S, \phi)}. \tag{13}$$

Then by equation (12) and by property (5) of evidence (Subsection 2.1),

$$m = \frac{E(S \cup \{\phi\}, \neg\psi)}{E(S, \neg\psi)} = \frac{1 - E(S \cup \{\phi\}, \psi)}{1 - E(S, \psi)} = \frac{1 - kE(S, \psi)}{1 - E(S, \psi)}. \tag{14}$$

So, $m < 1$ ($m > 1$ or $m = 1$) iff $k > 1$ ($k < 1$ or $k = 1$, respectively).

(c) By equations (12–14),

$$\delta_\phi(\psi) = (k-1)E(S, \phi); \quad \delta_\phi(\neg\psi) = (1-k)\frac{E(S, \psi)}{1 - E(S, \psi)}E(S, \phi).$$

So, for a given value of $E(S, \psi)$, $|\delta_\phi(\psi)|$ and $|\delta_\phi(\neg\psi)|$ differ by a constant factor. ■

So, considering an explanation $\phi$, if there exists a testable formula $\theta$ not indifferent to $\phi$, then measuring the value of $\theta$ may help in discriminating between explanations by increasing or decreasing the credibility of $\phi$. The larger $|\delta_\phi(\theta)|$ and $|\delta_\phi(\neg\theta)|$, the stronger the discriminating ability of $\theta$.

**Definition 5.2** *A testable formula $\theta \in TF$ is a* discriminating test *for an explanation $\phi$ if*

$$|\delta_\phi(\theta)| + |\delta_\phi(\neg\theta)| = \max_{\psi \in TF} \left[|\delta_\phi(\psi)| + |\delta_\phi(\neg\psi)|\right].$$

Let us consider again the malfunctioning adder of Section 3 (figure 1), and assume that it is possible to measure signals in nodes $u, v, w$ of the circuit, so, $TF = \{u, v, w\}$. Table 7 presents, for all diagnoses $\Delta_i$, $i = 1, \ldots, 6$ (equation 8) and all testable formulas $\theta \in TF$, values of $\delta 0 = \delta_{\Delta_i}(\theta = 0)$, $\delta 1 = \delta_{\Delta_i}(\theta = 1)$, $\delta 2 = |\delta 0| + |\delta 1|$. For each diagnosis a maximum value of $\delta 2$ is shown in bold face.

**Table 7**

| $i$ | $u$ | | | $v$ | | | $w$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\delta 0$ | $\delta 1$ | $\delta 2$ | $\delta 0$ | $\delta 1$ | $\delta 2$ | $\delta 0$ | $\delta 1$ | $\delta 2$ |
| 1 | 0.160 | -0.160 | **0.320** | 0.081 | -0.107 | 0.188 | 0.081 | -0.107 | 0.188 |
| 2 | 0.107 | -0.107 | 0.214 | 0.054 | -0.071 | 0.125 | -0.321 | 0.429 | **0.750** |
| 3 | 0.107 | -0.107 | 0.214 | -0.321 | 0.429 | **0.750** | 0.054 | -0.071 | 0.125 |
| 4 | -0.321 | 0.321 | **0.642** | 0.054 | -0.071 | 0.125 | 0.054 | -0.071 | 0.125 |
| 5 | -0.214 | 0.214 | 0.428 | 0.036 | -0.048 | 0.084 | -0.214 | 0.286 | **0.500** |
| 6 | -0.214 | 0.214 | 0.428 | -0.214 | 0.286 | **0.500** | 0.036 | -0.048 | 0.084 |

So, measuring $u$ presents a discriminating test for both $\Delta_1$ and $\Delta_4$, such that if $u = 0$ then the evidence of $\Delta_1$ increases from 0.482 (table 3) to 0.642, while the evidence of $\Delta_4$ decreases from 0.321 to zero meaning that $\Delta_4$ is not an explanation any more. However, if $u = 1$ then the evidence of $\Delta_1$ drops to 0.322, while the evidence of $\Delta_4$ rises to 0.642 such that $\Delta_4$ becomes the most likely diagnosis. Similarly, measuring $v$ presents a discriminating test for $\Delta_3$ and $\Delta_6$.

If the evidence of an explanation $\phi$ raised by a test $\theta$ is still not sufficiently high (e.g., 0.642 for $\Delta_1$ in the example above), then the measured value of $\theta$ should be added to the original observation, and then a new discriminating test for $\phi$ should be performed under the extended observation $Obs' = Obs \cup \{\theta\}$.

Thus, reasoning by evidence suggests the following procedure for determining a most likely explanation.

**Procedure Best Explanation**

Given a system description $S$, an observation $Obs$ and a set of possible explanations $EXPL$, Repeat

{ Determine a currently best explanation $\phi \in EXPL$ such that

$$E(S \cup Obs, \phi) = \max_{expl \in EXPL} [E(S \cup Obs, expl)] ;$$

If the value of $E(S \cup Obs, \phi)$ is not satisfactorily high, then do

{ Determine a discriminating test $\theta$ for $\phi$;

Measure the value of $\theta$;

Augment the current observation $Obs := Obs \cup \{\theta\}$;

Update $EXPL$

}

}

Until the best explanation is sufficiently credible or there is no discriminating test available. ■

The same procedure is applicable to determining of a best conjecture.

# 6 Counting models

Reasoning by evidence requires counting models of logic formulas which is a hard computational task. Valiant [85] has shown that in general the problem is #P-complete. Due to Roth [73], counting models of propositional languages is exponentially hard even for Horn and monotone formulas. Dubois [17] and Zhang [88] have developed algorithms for computing the number of models of a propositional CNF formula (with $n$ variables, $m$ clauses, $k$ literals in each clause) having a worst case time complexity of $O(m\alpha_k^n)$, where $\alpha_k$ is the largest root of $x^k - x^{k-1} - \ldots - x - 1$ (such as $\alpha_2 = 1.618$, $\alpha_3 = 1.839$, $\alpha_4 = 1.928, \ldots, \lim_{k\to\infty} \alpha_k = 2$).

Very hard in general, the problem of counting models turns out to be quite tractable in many practical cases. Luby and Velickovic [53] have presented an algorithm for approximating the number of models of a DNF propositional formula with $m$ clauses on $n$ variables. The running time of the algorithm is estimated by a polynomial in $m$ and $n$ multiplied by $n^{O(\frac{1}{\delta}(\log^2 m \,+\, \log^2 \frac{1}{\epsilon})(\log\log m \,+\, \log\frac{1}{\epsilon} \,+\, \log\frac{1}{\delta}))}$, where $\epsilon$, $\delta$ are the approximation errors. Linial and Nisan [47] have presented an algorithm for computing the size of a union of a family of sets, that can be used for approximating the number of models. The algorithm runs in time $O(m^r)$, where $r$ grows with the precision. It has been shown in [49] that under reasonable assumptions the average run time of computing the exact number of models of a propositional formula is $O(m^c n)$ where $c = O(\log m)$.

## 7 Conclusion

Given a system description $S$ and its observed behaviour $Obs$, there usually are several possible explanations of the observation constituting a set $EXPL$. Different explanations may be contradicting and suggesting conflicting actions. So, in order to take proper actions sanctioned by the real situation, we need to single out an explanation $\phi \in EXPL$ that is most adequate to the reality. Many existing methods for determining of such a *best explanation* require an extensive statistical knowledge about the system behaviour which often is either only partially available or not satisfactorily reliable.

Trying to overcome this lack of data, we introduce the notion of *evidence* as a measure of credibility of beliefs based on the semantic information contained in the system and supplied by the observations. Then evidence is applied to the problem of determining a best explanation.

Although it is most likely that a best explanation conforms to the reality, there still is a chance that some other possible explanation is the actual one. To increase our confidence in the preferred explanation and to discredit other possible ones we need to acquire more information about the present situation by means of performing *tests*. The notion of evidence turns out to be useful in determining tests for discrimination among possible explanations.

It may happen that the available knowledge and accessibility of the system do not enable determining an explanation that fully describes the actual causes of the observed situation. In such cases a reasonable *conjecture* about the state of the system would be helpful. It is shown how evidence can be used for determining a most likely conjecture.

As calculating evidence is based on counting models of the system, we address briefly this problem indicating that in many practical cases counting propositional models is tractable. On the other hand, counting models of first order formulas remains an open problem.

## References


[1] Aczel, J., and Forte, B., 1986, Generalized Entropies and the Maximum Entropy Principle. In *Maximum Entropy and Bayesian Methods in Applied Statistics*. Ed. by J. H. Justice (Cambridge University Press) 95-100.

[2] Andreassen, S., Woldbye, M., Falk, B., and Andersen, S., 1987, MUNIN – A causal probabilistic network for interpretation of electromyographic findings. In *Proceedings, 10th International Joint Conference on Artificial Intelligence*, Milano, 366–372.

[3] Besnard, P., and Cordier, M. O., 1992, Explanatory diagnoses and their computation by circumscription. In *Proceedings, 10th European Conference on Artificial Intelligence*, 724–728.

[4] Böttcher, C., 1995, No faults in structure? How to diagnose hidden interactions. In *Proceedings, International Joint Conference on Artificial Intelligence*, 1728–1733.

[5] Buchler, J. (Ed.), 1955, *Philosophical Writings of C. S. Peirce*, Dover, N. Y.

[6] Chajewska, U., and Halpern, J. Y., 1997, Defining explanation in probabilistic systems. In *Proceedings, 13th Conference on Uncertainty in Artificial Intelligence*, D. Geiger and P. Shenoy, Eds., Providence, RI, 62–71.

[7] Chapman, D., 1987, Planning for conjunctive goals. *Artificial Intelligence* **32**: 333–356.

[8] Charniak, E., and Santos, E. Jr., 1992, Dynamic MAP calculations for abduction. In *Proceedings, 10th Natl. Conf. on Artificial Intelligence, AAAI-92*, 552–557.

[9] Charniak, E., and Shimony, S. E., 1994, Cost-based abduction and MAP explanation. *Artificial Intelligence*, **66**:345–374.

[10] Cook, S., 1971, The complexity of theorem-proving procedures. In *Proceedings, 3rd ACM Symp. on Theory of Computing*, Shaker Heights, Ohio, ACM Press, 151–158.

[11] Darwiche, A., 1995, Model-based diagnosis using causal networks. In *Proceedings, International Joint Conference on Artificial Intelligence*, 211–217.

[12] Davis, M., and Putnam, H., 1960, A computing procedure for quantification theory. *J. ACM*, v. 7, no. 3, 201-215.

[13] Davis, R., 1984, Diagnostic reasoning based on structure and behaviour. *Artificial Intelligence* **24**: 347–410.

[14] DeJong, G., 1990, Plausible inference vs. abduction. *Proc. AAAI Symposium on Automated Abduction*, Stanford, 48–51.

[15] Dempster, A. P., 1967, Upper and lower probabilities induced by a multivariate mapping. *Annals of Mathematical Statistics*, **38**: 325–339.

[16] Dempster, A. P., 1968, A generalization of Bayesian inference. *Journal of the Royal Statistical Society*, **30**: 205–247.

[17] Dubois, O., 1991, Counting the number of solutions for instances of satisfiability. *Theoretical Computer Science*, **81**: 49–64.

[18] Eshghi, K., 1988, Abductive planning with event calculus. In *Proceedings, 5th International Conference and Symposium on Logic Programming*, 562–579.

[19] Eshghi, K., 1993, A tractable class of abduction problems. In *Proceedings, 13th International Joint Conference on Artificial Intelligence*, Chambery, France, 3–8.

[20] Fink, P., Lusth, J., and Duran, J., 1985, A general expert system design for diagnostic problem solving. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 553–560.

[21] Friedrich, G., Gottlob, G., and Nejdl, W., 1992, Formalizing the repair process. In *Proceedings, 10th European Conference on Artificial Intelligence*, B. Neumann (Ed.), 709–713.

[22] Gallo, G., and Urbani, G., 1989, Algorithms for testing the satisfiability of propositional formulae. *J. Logic Programming*, **7**: 45–61.

[23] Gärdenfors, P., 1988, *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, MIT Press.

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

[25] Genesereth, M. R., and Nilsson, N. J., 1987, *Logical Foundations of Artificial Intelligence*, Morgan Kaufmann, Los Altos, CA.

[26] Goldberg, A., 1979, Average case complexity of the satisfiability problem. In *Proceedings, 4th Workshop on Automated Deduction*, Austin, TX, 1–6.

[27] Goldberg, A., Purdom, P., and Brown, C., 1982, Average time analysis of simplified Davis-Putnam procedures. *Information Processing Letters*, **15**: 72–75.

[28] Goldszmidt, M., Morris, P., and Pearl, J., 1990, A maximum entropy approach to non-monotonic reasoning. In *Proceedings, 8th Natl. Conf. on Artificial Intelligence, AAAI-90*, Boston, Mass., The MIT Press, Menlo Park, CA, 646–652.

[29] Grove, A. J., Halpern, J. Y., and Koller, D., 1994, Random worlds and maximum entropy. *Journal of Artificial Intelligence Research*, **2**: 33–88.

[30] Hartley, R. V., 1928, Transmission of information. *Bell Systems Technical Journal*, **7**: 535–563.

[31] Heckerman, D., 1986, Probabilistic interpretation of MYCIN's certainty factors. In *Uncertainty in Artificial Intelligence*. Ed. by L. N. Kanal and J. F. Lemmer (North Holland) 167–196.

[32] Hobbs, J. R., Stickel, M. E., Martin, P., and Edwards, D., 1988, Interpretation as abduction. In *Proceedings, 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo, N. Y., 95–103.

[33] Hobbs, J. R., 1990, An integrated abductive framework for discourse interpretation. In *Proceedings, AAAI Symposium on Automated Abduction*, Stanford, 10–12.

[34] Iwama, K., 1989, CNF satisfiability test by counting and polynomial average time. *SIAM Journal on Computing*, **18**: 385–391.

[35] Jaynes, E. T., 1957, Information Theory and Statistical Mechanics. *Physical Review*, **106**: 620–630.

[36] Jaynes, E. T., 1979, Where do we stand on maximum entropy? In *The Maximum Entropy Formalism*, Ed. by R. D. Levine and M. Tribus (Cambridge, MA: MIT Press) 15–118.

[37] Jaynes, E. T., 1985, Where do we go from here? In *Maximum-Entropy and Bayesian Methods in Inverse Problems*, Ed. by C. R. Smith and W. T. Grandy (Reidel Publishing Co., Dordrecht, Holland) 21–58.

[38] Justice, J. H., Ed., 1986, *Maximum Entropy and Bayesian Methods in Applied Statistics* (Cambridge University Press).

[39] Kass, A., and Leake, D., 1988, Case-based reasoning applied to constructing explanations. In *Proceedings, Case-Based Reasoning Workshop*, J. Kolodner (Ed.), Palo Alto, 190–208.

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

[41] de Kleer, J., Mackworth, A. K., and Reiter, R., 1990, Characterizing diagnoses. In *Proceedings, 8th Natl. Conf. on Artificial Intelligence, AAAI-90*, Boston, Mass., The MIT Press, Menlo Park, CA, 324–330.

[42] Konolige, K., 1994, Using default and causal reasoning in diagnosis. *Annals of Mathematics and Artificial Intelligence* **11**: 97–135.

[43] Leake, D. B., 1992, *Evaluating Explanations: A Content Theory.* Lawrence Erlbaum, Hillsdale, N. J..

[44] Leake, D. B., 1993, Focusing construction and selection of abductive hypotheses. In *Proceedings, 13th International Joint Conference on Artificial Intelligence*, Chambery, France, 24–29.

[45] Lee, N. S., Grize, Y. L., and Dehnad, K., 1987, Quantitative models for reasoning under uncertainty in Knowledge-Based Expert Systems. *International Journal of Intelligent Systems*, **2**: 15–38.

[46] Levesque, H. J., 1989, A knowledge-level account of abduction. In *Proceedings, 11th International Joint Conference on Artificial Intelligence*, Detroit, Michigan, 1061–1067.

[47] Linial, N., and Nisan, N., 1990, Approximate inclusion-exclusion. *Combinatorica*, **10**(4): 349–365.

[48] Lipton, P., 1991, *Inference to the best explanation* (London: Routledge).

[49] Lozinskii, E., 1992, Computing propositional models. *Information Processing Letters*, **41**: 327–332.

[50] Lozinskii, E., 1993, A simple test improves checking satisfiability. *Journal of Logic Programming*, **15**: 99–111.

[51] Lozinskii, E., 1994, Information and evidence in logic systems. *Journal of Experimental and Theoretical Artificial Intelligence*, **6**: 163–193.

[52] Lozinskii, E., 1995, Is there an alternative to parsimonious semantics? *Journal of Experimental and Theoretical Artificial Intelligence*, **7**: 361–378.

[53] Luby, M., and Velickovic, B., 1991, On deterministic approximation of DNF. In *Proceedings, 23rd ACM Symp. on Theory of Computing*, New Orleans, LA, ACM Press, 430–438.

[54] McCarthy, J., 1980, Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, **13**: 27–39.

[55] McIlraith, S., 1994, Further contributions to characterizing diagnosis. *Annals of Mathematics and Artificial Intelligence*, **11**: 137–167.

[56] McIlraith, S., 1994, Generating tests using abduction. In *Proceedings, 1st International Conference on Principles of Knowledge Representation and Reasoning*, Bonn, Germany, 449–460.

[57] Miller, R., Pople, H., and Myers, J., 1982, INTERNIST-1, An experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 468–476.

[58] Mozetič, I., 1992, A polynomial-time algorithm for model-based diagnosis. In *Proceedings, 10th European Conference on Artificial Intelligence*, 729–733.

[59] Ng, H. T., and Mooney, R. J., 1992, Abductive plan recognition and diagnosis: a comprehensive empirical evaluation. In *Proceedings, 3rd International Conference on Principles of Knowledge Representation and Reasoning*, Cambridge, 499–508.

[60] Nilsson, N. J., 1986, Probabilistic logic. *Artificial Intelligence*, **28**: 71–87.

[61] O'Rorke, P., 1990, An approach to theory revision using abduction. In *Proceedings, AAAI Symposium on Automated Abduction*, Stanford, 33–37.

[62] Patil, R., 1981, Causal representation of patient illness for electrolyte and acid-base diagnosis. Technical Report MIT/LCS/TR–267, Cambridge, Mass.

[63] Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann, Los Altos, CA).

[64] Pearl, J., 1989, Probabilistic semantics for nonmonotonic reasoning: a survey. In *Proceedings, First Intl. Conf. on Principles of Knowledge Representation and Reasoning*, Toronto, Morgan Kaufmann, San Mateo, CA, 699–710.

[65] Peng, Y., and Reggia, J. A., 1987, A probabilistic causal model for diagnostic problem solving. *IEEE Transactions on Systems*, vol. SMC-17, no. 2, 146–162.

[66] Poole, D., 1991, Representing diagnostic knowledge for probabilistic Horn abduction. In *Proceedings, 12th International Joint Conference on Artificial Intelligence*, Sydney, 1129–1135.

[67] Poole, D., and Provan, G. M., 1991, What is the most likely diagnosis? In: *Uncertainty in Artificial Intelligence*, P. P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer (Eds.) Elsevier, 89–105.

[68] Pople, H., 1982, Heuristic methods for imposing structure on ill-structured problems: The structuring of medical diagnostics. *Artificial Intelligence in Medicine*, P. Szolovits (Ed.) 119–190.

[69] Purdom, P., and Brown, C., 1985, The pure literal rule and polynomial average time. *SIAM Journal on Computing*, **14**(4): 943–953.

[70] Reggia, J. A., Nau, D. S., and Wang, P., 1983, Diagnostic expert systems based on a set covering model. *International Journal of Man-Machine Study*, 437–460.

[71] Reiter, R., 1980, A logic for default reasoning. *Artificial Intelligence*, **13**: 81–132.

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

[73] Roth, D., 1996, On the hardness of approximate reasoning. *Artificial Intelligence*, **82**: 273–302.

[74] Schank, R. C., and Leake, D., 1989, Creativity and learning in a case-based explainer. *Artificial Intelligence*, **40**: 353–385.

[75] Schank, R. C., 1986, *Explanation Patterns: Understanding Mechanically and Creatively.* Lawrence Erlbaum, Hillsdale, N. J.

[76] Selman, B., and Levesque, H. J., 1990, Abductive and default reasoning: a computational core. In *Proceedings, 8th Natl. Conf. on Artificial Intelligence, AAAI-90*, Boston, Mass., The MIT Press, Menlo Park, CA, 343–348.

[77] Shafer, G., 1976, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, NJ).

[78] Shannon, C. E., 1948, A mathematical theory of communication. *Bell Systems Technical Journal*, **27**: 379–423, 623–656.

[79] Shimony, S. E., 1991, Explanation, irrelevance and statistical independence. In *Proceedings, 9th Natl. Conf. on Artificial Intelligence, AAAI-91*, Anaheim, CA, 482–487.

[80] Shortliffe, E. H., 1976, *Computer-Based Medical Consultation: MYCIN* (Elsevier).

[81] Skyrms, B., 1985, Maximum entropy inference as a special case of conditionalisation. *Synthese*, **63**: 55–74.

[82] Sy, B. K., 1992, Reasoning MPE to multiply-connected belief networks using message passing. In *Proceedings, 10th Natl. Conf. on Artificial Intelligence, AAAI-92*, 570–576.

[83] Szolovits, P., and Pauker, S., 1978, Categorical and probabilistic reasoning in medical diagnosis. Artificial Intelligence, **2**: 115–144.

[84] Tribus, M., 1961, Information Theory as the basis for Thermostatics and Thermodynamics. *Journal of Applied Mechanics*, **28**: 1–8.

[85] Valiant, L., 1979, The complexity of computing the permanent. *Theoretical Computer Science*, **8**: 189–201.

[86] Wiener, N., 1948, *Cybernetics, or Control and Communication in the Animal and the Machine* (Wiley, New York).

[87] Williams, P. M., 1980, Bayesian conditionalisation and the principle of minimum information. *British Journal for the Philosophy of Science*, **31**: 131–144.

[88] Zhang, W., 1996, Number of models and satisfiability of sets of clauses. *Theoretical Computer Science*, **155**: 277–288.