Logique & Analyse **165–166** (1999), 161–180

A FORMAL ANALYSIS OF DIAGNOSIS AND DIAGNOSTIC REASONING*

ERIK WEBER & DAGMAR PROVIJN

1. *Introduction*

Diagnostic reasoning may relate to an *established fault* in a *system* or an *established fault* in an *individual*. A system is to be understood as a structured whole of components, while an individual is an object that is not analysed into components.

With respect to systems, three types of diagnosis can be distinguished: non-explanatory, weak explanatory and strong explanatory. In section 2 we define these types, provide illustrations, and describe their respective functions. In section 3 we analyse the reasoning process by which non-explanatory diagnoses are constructed, and argue that the adaptive logics $\mathrm{AL}_{\mathrm{EXP}}$ and $\mathrm{AL}^*_{\mathrm{EXP}}$ are adequate tools for modelling this kind of diagnostic reasoning. In section 4 we discuss (weak and strong) explanatory diagnostic reasoning, and show that it can be divided in three stages. Each stage must be modelled by means of a different adaptive logic.

In section 5 we discuss diagnosis and diagnostic reasoning in individuals. We show that non-explanatory diagnoses do not occur here, while the conclusions of 4 can be extended to individuals.

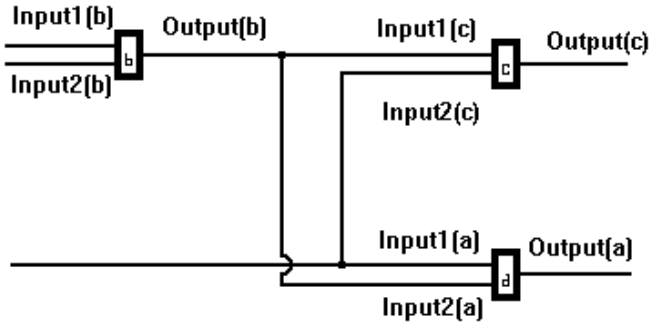
*We thank the members of the Centre for Logic and Philosophy of Science (Ghent University), especially Diderik Batens and Joke Meheus, for the comments on earlier versions of this paper. The research for this paper was supported by the Fund for Scientific Research-Flanders through research project G.0131.01 (Philosophical and technical foundations of the adaptive logic programme, further incorporation of logical mechanisms and further development of systems and applications), the Research Fund of Ghent University through research project BOF2001/GOA/008 ("Development of Adaptive Logics for the Study of Central Topics in Contemporary Philosophy of Science") and indirectly the Flemish Minister responsible for Science and technology (contract BIL 98/73).

2. *Non-explanatory and Explanatory Diagnoses for Faults in Systems*

2.1 Following Reiter 1987 (p. 59), we define systems as follows:

(2.1) A *system* is a pair (SD, COMP) where:
(a) SD, the system description, is a set of first-order-sentences
(b) COMP, the system components, is a finite set of constants.

We will require (and this differs from how Reiter specifies SD) that the system description contains (i) a description of the input processing behaviour of every component (a description of how inputs are transformed into outputs), and (ii) a description of the relations between the components. As an example, consider the following electric circuit, which contains three components (the three gates a, b and c):



A possible system description is:

Description of input processing behaviour
a is an AND-gate, i.e. output$(a) = 1$ iff input$_1(a) =$ input$_2(a) = 1$.
b is an XOR-gate, i.e. output$(b) = 1$ iff input$_1(b) \neq$ input$_2(b)$.
c is an XOR-gate, i.e. output$(c) = 1$ iff input$_1(c) \neq$ input$_2(c)$.

Relations between components
Output$(b) =$ input$_2(a)$.
Output$(b) =$ input$_1(c)$.
Input$_1(a) =$ input$_2(c)$.

2.2 An established fault in a system is defined as follows:

(2.2) SD $\cup$ OBS is an *established fault* in a system (SD, COMP) if and only if

(a) OBS is a set of first-order sentences describing the observed states of (some or all inputs and outputs of the system components), and
(b) SD $\cup$ OBS is inconsistent.

As illustration, assume that in our example we observe that

output$(c) = 1$ & output$(a) = 0$

while the inputs are

input$_1(b) = 1$ & input$_2(b) = 0$ & input$_1(a) = 1$.

The conjunction of the three input values and two output values with the SD is inconsistent, because from SD and the observed inputs we can derive that

output$(c) = 0$ & output$(a) = 1$

So we have an established fault in our system.

2.3 A non-explanatory diagnosis for an established fault in a system can be defined as follows:

(2.3) Ω is a *non-explanatory diagnosis* for a fault F if and only if
(a) it is of the form $\wedge\neg \mathrm{P_i}\alpha_\mathrm{i}$ (where α_i is a system-component and $\mathrm{P_i}$ a predicate describing the input processing behaviour of the system-component),
(b) SD contains the set $\Gamma = \{\mathrm{P}_1\alpha_1, ..., \mathrm{P_n}\alpha_\mathrm{n}\}$ (where $\mathrm{P}_1\alpha_1, ..., \mathrm{P_n}\alpha_\mathrm{n}$ are the formulas whose negations are the conjuncts of $\wedge\neg \mathrm{P_i}\alpha_\mathrm{i}$),
(c) $(\mathrm{SD}\backslash\Gamma)\cup\Delta\cup\mathrm{OBS}$ is consistent (where $\Delta = \{\neg\mathrm{P}_1\alpha_1, ..., \neg\mathrm{P_n}\alpha_\mathrm{n}\}$),
(d) for all proper subsets Γ' and Δ' of Γ and Δ, $(\mathrm{SD}\backslash\Gamma') \cup \Delta' \cup \mathrm{OBS}$ remains inconsistent, and
(e) every set Δ'' satisfying conditions (b)–(d) has at least as many elements as Δ.

In our example there is one non-explanatory diagnosis: "b is not an XOR-gate". If we define *potential* explanatory diagnoses as statements satisfying conditions (a) and (b) of 2.3, then there are seven such candidates:

a is not an AND-gate
b is not an XOR-gate
c is not an XOR-gate

(a is not an AND-gate) $\wedge$ (b is not an XOR-gate)
(a is not an AND-gate) $\wedge$ (c is not an XOR-gate)
(b is not an XOR-gate) $\wedge$ (c is not an XOR-gate)
(a is not an AND-gate) $\wedge$ (b is not an XOR-gate) $\wedge$ (c is not an XOR-gate)

The first and third possibility do not restore consistency, so they violate condition (c). The fourth, sixth and seventh possibility violate condition (d). Finally, condition (e) eliminates the fifth possibility.

Before turning to explanatory diagnoses, it is useful to point out the underlying ideas of definition 2.3:
(1) The two parts of the system description SD (input processing behaviour and relation between components) have a different epistemological status: the claims about relations between components is not doubted, even if a fault is established. The claims about input processing behaviour can be given up if a fault is established; they ought to be interpreted as hypotheses which are believed to be true, but nonetheless are falsifiable through experimental observation. The observations OBS have the same epistemological status as the claims about the relations between components.
(2) The sole aim of diagnosis of this type is to restore consistency: the empirical data OBS are not explained by it. Therefore we call this type of diagnosis non-explanatory.
(3) The aim of non-explanatory diagnosis is to restore consistency in a parsimonious way; which is expressed in condition (d) and (e).

2.4 Weak explanatory diagnosis is defined as follows:

(2.4) If $\wedge\neg \mathrm{P_i}\alpha_\mathrm{i}$ is a *non-explanatory* diagnosis for a fault in a system, then $\wedge \mathrm{Q_i}\alpha_\mathrm{i}$ (where $\mathrm{Q_i}$ is also a predicate describing the input processing behaviour) is a *weak explanatory* diagnosis for the same fault if and only if
(a) $\wedge \mathrm{Q_i}\alpha_\mathrm{i}$ entails $\wedge\neg \mathrm{P_i}\alpha_\mathrm{i}$, and
(b) $(\mathrm{SD}\backslash\Gamma) \cup \mathrm{E}$ entails OBS (where $\Gamma = \{\mathrm{P}_1\alpha_1, ..., \mathrm{P_n}\alpha_\mathrm{n}\}$ and where $\mathrm{E} = \{\mathrm{Q}_1\alpha_1, ..., \mathrm{Q_n}\alpha_\mathrm{n}\}$).

Diagnoses of this type are called explanatory because the empirical data OBS are explained by them. The definition entails that multiple weak explanatory diagnoses can coexist.

2.5 Strong explanatory diagnoses are defined in such a way that they are unique:

(2.5) If $\wedge\neg P_i\alpha_i$ is a *non-explanatory* diagnosis for a fault in a system, then $\wedge Q_i\alpha_i$ (where Q_i is also a predicate describing the input processing behaviour) is a *strong explanatory* diagnosis for the same fault if and only if
(a) $\wedge Q_i\alpha_i$ entails $\wedge\neg P_i\alpha_i$,
(b) $(SD\backslash\Gamma) \cup E$ entails OBS, and
(c) for all sets $F = \{R_1\alpha_1, ..., R_n\alpha_n\}$ different from E, $(SD\backslash\Gamma) \cup F$ is incompatible with OBS.

Γ and E have the same meaning as above. $R_1\alpha_1, ..., R_n\alpha_n$ describe input processing behaviour.

In our example there are eight weak explanatory diagnoses. The gates in our circuit have one output and two inputs. Gates of this kind can be divided into 16 types, depending on their input processing behaviour:

T_1 1 1 1 1 0 1 0 1 1 0 0 1 TAUT	T_2 1 1 1 1 0 1 0 1 1 0 0 0 OR	T_3 1 1 1 1 0 1 0 1 0 0 0 1	T_4 1 1 1 1 0 0 0 1 1 0 0 1 IMPL
T_5 1 1 0 1 0 1 0 1 1 0 0 1 NOT-AND	T_6 1 1 1 1 0 1 0 1 0 0 0 0 LEFT	T_7 1 1 1 1 0 0 0 1 1 0 0 0 RIGHT	T_8 1 1 1 1 0 0 0 1 0 0 0 1 EQ
T_9 1 1 0 1 0 1 0 1 1 0 0 0 XOR	T_{10} 1 1 0 1 0 1 0 1 0 0 0 1 NOT-RIGHT	T_{11} 1 1 0 1 0 0 0 1 1 0 0 1 NOT-LEFT	T_{12} 1 1 1 1 0 0 0 1 0 0 0 0 AND
T_{13} 1 1 0 1 0 1 0 1 0 0 0 0 NOT-IMPL	T_{14} 1 1 0 1 0 0 0 1 1 0 0 0	T_{15} 1 1 0 1 0 0 0 1 0 0 0 1 NEITHER	T_{16} 1 1 0 1 0 0 0 1 0 0 0 0 CONTR

Starting from the non-explanatory diagnosis "b is not an XOR-gate", we obtain the following weak explanatory diagnoses:

b is an IMPL-gate	b is a RIGHT-gate
b is an EQ-gate	b is a NOT-LEFT-gate
b is an AND-gate	b is T_{14}-gate
b is a NEITHER-gate	b is a CONTR-gate

Combined with (SD\Γ), each of these eight possibilities yields a new system description that explains the observations the inquirer has made.

The fact that we have eight weak explanatory diagnoses in our example entails that there is no strong one. A strong diagnosis can be obtained by asking questions whose answers eliminate some of the weak diagnoses. For instance, we can ask what happens if the inputs are changed into:

$$\text{input}_1(b) = 1 \;\&\; \text{input}_2(b) = 1 \;\&\; \text{input}_1(a) = 1.$$

Let us assume that for these inputs the same outputs are observed as in the original observation:

$$\text{output}(c) = 1 \;\&\; \text{output}(a) = 0$$

If we compare the eight possibilities with the results of this measurement, we discover that four of them are falsified: if b is an IMPL-gate, a RIGHT-gate, an AND-gate or an EQ-gate, then $\text{output}(b) = 1$, and thus $\text{output}(c) = 0$ and $\text{output}(a) = 1$. This contradicts the observations. On the other hand, if b is a NOT-LEFT-gate, a T_{14}-gate, a NEITHER-gate or a CONTR-gate, then $\text{output}(b) = 0$, and thus $\text{output}(c) = 1$ and $\text{output}(a) = 0$. This is what we have observed, so these four possible explanations remain. By performing more measurements and making the corresponding calculations, we can try to exclude all possibilities but one. We will not always succeed in doing this (see section 4 for details).

2.6 Each type of diagnosis has one or more characteristic functions. The original system description usually describes the way in which the system is designed to behave. Non-explanatory diagnoses can help us to *repair* faults in a system: the component that is diagnosed to behave differently from what is expected, can be replaced. The aim of repairing is to ensure that the system behaves as described in the original system description. This is the primary function of non-explanatory diagnoses.

Weak and strong explanatory diagnoses give us *more options* for repairing the system. If we have a well-supported hypothesis about what is wrong with a component (rather than a mere non-explanatory diagnosis), we can try to compensate the fault by changes in other components or in the relations between the components. In the case of strong diagnoses, the result of these changes can be reliably predicted; in the case of a weak diagnosis, this is

often impossible (because the different diagnoses would lead to different outcomes if the same change outside the component is made).

Besides these primary functions (which relate to the intrinsic uses of the diagnoses), we can distinguish secondary functions: non-explanatory diagnoses are useful steps for constructing weak explanatory diagnoses, and weak explanatory diagnoses help us to construct strong ones. These relations will be further analysed in section 4.

3. *Logical Analysis of Non-explanatory Diagnostic Reasoning in Systems*

3.1 The aim of this section is to show that non-explanatory diagnostic reasoning can be adequately modelled by means of the adaptive logics $\mathrm{AL}_{\mathrm{EXP}}$ and $\mathrm{AL}^*_{\mathrm{EXP}}$. These logics will be presented in 3.2–3.4, while their application is clarified in sections 3.5 and 3.6. Some preliminary remarks are necessary:
(1) The epistemological difference between the two parts of SD (cfr. section 2.3) is incorporated in both logics by distinguishing *premises that are not doubted* and *expected premises (expectations)*.
(2) It is assumed that the *premises that are not doubted* (i.e. observations and claims about relations between components) are consistent. Inconsistencies can arise only when the latter are used in combination with *expectations*.
(3) The derivability relation of both logics is ampliative: the CL-consequences (where CL stands for the propositional part of Classical Logic) of a set of premises are a proper subset of the $\mathrm{AL}_{\mathrm{EXP}}$- and the $\mathrm{AL}^*_{\mathrm{EXP}}$-consequences. The reason is that conclusions can be drawn from *expectations*, which is impossible in CL-proofs.
(4) What is derived from *expectations*, can only be derived *conditionally*. This means that remains derivable as long as the *expectations* on which it depends do not lead to triviality.
(5) Consequently, $\mathrm{AL}_{\mathrm{EXP}}$- and $\mathrm{AL}^*_{\mathrm{EXP}}$-proofs are dynamic. As long as there are no inconsistencies, all CL-rules can be used to derive conclusions from *expectations*. But as soon as the proof contains an inconsistency, some of the formulas that are derived conditionally are marked in the proof. As long as these formulas are marked, they are considered as not being derivable from the given set of premises. $\mathrm{AL}_{\mathrm{EXP}}$ and $\mathrm{AL}^*_{\mathrm{EXP}}$ share this dynamic proof format with all other adaptive logics (see Batens 1998 and 2000, and Weber & De Clercq 200+).
(6) The formal difference between $\mathrm{AL}_{\mathrm{EXP}}$ and $\mathrm{AL}^*_{\mathrm{EXP}}$ lies in the conditions (expressed in two different marking definitions) under which formulas must be marked. In $\mathrm{AL}_{\mathrm{EXP}}$ more lines are marked than in $\mathrm{AL}^*_{\mathrm{EXP}}$. As we will see in 3.5 and 3.6, this rather small formal difference causes a big difference in

the philosophical significance of the logics.

3.2 An adaptive logic oscillates between an upper- and lower limit logic. The lower limit logic defines the 'safe' derivations; hence, its inference rules can be applied unconditionally. The upper limit logic will decide on what we believe to be 'normal'. What counts as normal will be determined by the kind of reasoning we want to capture. Within the context of non-explanatory diagnostic reasoning, CL will be the lower limit logic, while CL_{EXP} is the upper limit logic. CL_{EXP} presupposes that all expectations are true.

The language scheme of CL_{EXP} is an extension of the CL-language scheme with the logical term E. The wff (well-formed formula) E(A) expresses the idea that wff A is expected. A will be called the factor of expectation E(A). The CL_{EXP} derivation rules are also an extension of the CL derivation rules. There is only one extra rule:

EXP: if E(A) occurs in a CL_{EXP}-proof, then add A to it.

This rule assures that A_1, ..., A_n, $E(B_1)$, ..., $E(B_m) \vdash_{CLexp} C$ if and only if A_1, ..., A_n, B_1, ..., $B_m \vdash_{CL} C$. In other words: CL_{EXP} presupposes that every *expectation* is consistent with the (other) premises. When A is a "bad" expectation, CL_{EXP} will generate trivial inferences from the given set of premises. The adaptive logics AL_{EXP} and AL^*_{EXP} have the same conclusions as CL_{EXP} when the premises and expectations are consistent, but avoid triviality in the inconsistent case.

3.3 AL_{EXP}- and AL^*_{EXP}-proofs are written in a specific format according to which each line of a proof consists of five elements:

> (i) a line number,
> (ii) the wff derived,
> (iii) the line numbers of the wffs from which (ii) is derived,
> (iv) the inference rule that justifies the derivation, and
> (v) the set of the line numbers of the expectations on which we rely in order for (ii) to be derivable by (iv) from the formulas on the lines enumerated in (iii).

In constructing AL_{EXP}- and AL^*_{EXP}-proofs, one is allowed to use a structural rule and three generic inference rules. The marking definitions of both adaptive logics (M and M*) must be taken into account at each new step of the proof: after a line has been introduced, we check whether the line itself or previous lines should be marked or unmarked.

STRUCTURAL RULE

Premise rule

PREM At any stage of a proof one may add a line consisting of (i) an appropriate line number, (ii) a premise, (iii) a dash, (iv) 'PREM', (v) '$\emptyset$'.

GENERIC INFERENCE RULES

Unconditional rule

RU If A_1, ..., $A_n \vdash_{CL} B$, and A_1, ..., A_n occur in the proof, then one may add B to the proof. The fifth element of the new line is the conjunction of the fifth elements of the lines in its third element.

Conditional rule

RC If E(A) occurs in an AL_{EXP}-proof (or AL^*_{EXP}-proof), then one may add A to it. The fifth element is the line number of the line which contains E(A) as its third element.

Derived rule

RD If $A\&{\sim}A$ is derivable under conditions $\{i\}$, with k, ..., $m \in \{i\}$ as the line numbers of the expectations $E(B_k)$, ..., $E(B_m)$; then one may add $(E(B_k)\&{\sim}B_k) \vee ... \vee (E(B_m)\&{\sim}B_m)$ to the proof with $\emptyset$ as its fifth element.

MARKING DEFINITIONS

Let $\text{DAB}(A_1, ..., A_n)$ refer to the formula $(E(A_1)\&{\sim}A_1) \vee ... \vee (E(A_n)\&{\sim}A_n)$, with A_1, ..., A_n as its factors. "DAB" stands for "disjunction of abnormalities". As it is normal that both E(A) and A occur in the proof, the presence of both E(A) and ${\sim}A$ is abnormal. Before we can give the marking definitions, we need a number of preliminary definitions.

Definition 1: DAB*-consequence*

$\text{DAB}(A_1, ..., A_n)$ is a DAB-consequence of Γ iff $\text{DAB}(A_1, ..., A_n)$ is CL-derivable from Γ.

Definition 2: Minimal DAB*-consequence at a stage of a proof*

A minimal DAB-consequence of Γ at a stage of a proof is a DAB-consequence of Γ derived at that stage of the proof such that no result of dropping a disjunct from it is derived at that stage of the proof.

Definition 3: Set of factors at a stage of a proof

$U_s(\Gamma) = \{A \mid A$ is a factor of a minimal DAB-consequence of Γ at stage s of the proof$\}$.

*Definition 3**

$U_s{}^*(\Gamma) = \cup\{\Sigma \mid \Sigma \in P\ U_s(\Gamma)$, with Σ containing a factor of each minimal DAB-consequence of Γ at stage s and there is no $\Sigma' \in P\ U_s(\Gamma)$, with Σ' containing a factor of each minimal DAB-consequence of Γ at stage s such that $\#\Sigma' < \#\Sigma\}$.

The marking definitions for both logics are:

Marking definition AL_{EXP}

M: a line is marked M iff where Δ is the set of factors of the expectations

denoted by its fifth element, $\Delta \cap \mathrm{U_s}(\Gamma) \neq 0$.
Marking definition AL$^*_{\mathrm{EXP}}$
M*: a line is marked M* iff where Δ is the set of factors of the expectations denoted by its fifth element, $\Delta \cap \mathrm{U_s}^*(\Gamma) \neq 0$.

3.4 The dynamic character of the proofs requires a distinction between derivability at a stage of the proof and final derivability.
Derivability at a stage of a proof
A is derived at a stage of a proof in a proof from Γ iff A is not marked at that stage of the proof.
Final derivability
A is finally derived in a proof from Γ iff A is derived at a line that is not marked and, any extension of the proof in which A is marked, may be further extended in such way that A becomes unmarked.

3.5 To clarify the applicability of both logics in the context of non-explanatory diagnosis, we will consider a proof in which "b is not an XOR-gate" is obtained as a non-explanatory diagnosis for the fault discussed in section 2. We will first apply AL$^*_{\mathrm{EXP}}$ on this example. We will use a propositional language with primitive sentences P^0, P^1, P^2, Q^0, Q^1, Q^2, R^0, R^1, R^2. P^0 means that the output-value of gate a is 1. P^1 means that input$_1$ of a has value 1. P^2 means that output$_2$ of a is 1. The same can be done for Q (gate b) and R (gate c).

1	$\mathrm{Q}^0 \equiv \mathrm{P}^2$	-	PREM	$\emptyset$
2	$\mathrm{Q}^0 \equiv \mathrm{R}^1$	-	PREM	$\emptyset$
3	$\mathrm{P}^1 \equiv \mathrm{R}^2$	-	PREM	$\emptyset$
4	$E(\mathrm{P}^0 \equiv (\mathrm{P}^1 \& \mathrm{P}^2))$	-	PREM	$\emptyset$
5	$E(\mathrm{Q}^0 \equiv\sim (\mathrm{Q}^1 \equiv \mathrm{Q}^2))$	-	PREM	$\emptyset$
6	$E(\mathrm{R}^0 \equiv\sim (\mathrm{R}^1 \equiv \mathrm{R}^2))$	-	PREM	$\emptyset$
7	R^0	-	PREM	$\emptyset$
8	$\sim \mathrm{P}^0$	-	PREM	$\emptyset$
9	Q^1	-	PREM	$\emptyset$
10	$\sim \mathrm{Q}^2$	-	PREM	$\emptyset$
11	P^1	-	PREM	$\emptyset$
12	$\mathrm{P}^0 \equiv (\mathrm{P}^1 \& \mathrm{P}^2)$	4	RC	{4}
13	$\mathrm{Q}^0 \equiv\sim (\mathrm{Q}^1 \equiv \mathrm{Q}^2)$	5	RC	{5}
14	$\mathrm{R}^0 \equiv\sim (\mathrm{R}^1 \equiv \mathrm{R}^2)$	6	RC	{6}
15	$\sim (\mathrm{Q}^1 \equiv \mathrm{Q}^2) \supset \mathrm{Q}^0$	13	RU	{5}
16	Q^0	9,10,15	RU	{5}
17	R^1	2,16	RU	{5}

18	R^2	3,11	RU	$\emptyset$
19	$R^0 \supset\sim (R^1 \equiv R^2)$	14	RU	{6}
20	$\sim R^0$	17,18,19	RU	{5,6}
21	P^2	1,16	RU	{5}
22	$(P^1\&P^2) \supset P^0$	12	RU	{4}
23	$P^1\&P^2$	11,21	RU	{5}
24	P^0	22,23	RU	{4,5}
25	$\sim R^0\&P^0$	20,24	RU	{4,5,6}
26	$\sim P^0\&P^0$	8,24	RU	{4,5,6}
27	$[E(P^0 \equiv (P^1\&P^2))\& \sim (P^0 \equiv (P^1\&P^2))]\vee$ $[E(Q^0 \equiv\sim (Q^1 \equiv Q^2))\& \sim (Q^0 \equiv\sim (Q^1 \equiv Q^2))]\vee$ $[E(R^0 \equiv\sim (R^1 \equiv R^2))\& \sim (R^0 \equiv\sim (R^1 \equiv R^2))]$	26	RD	$\emptyset$

At line 27 we derive the first DAB-consequence. Because the contradiction at line 26 has all three expectations as its condition, all conditional lines of the proof must be marked (i.e., 12 till 26, except 18). We now try to derive more DAB-consequences:

28	R^1	2,16	RU	{5}	M*
29	R^2	3,11	RU	$\emptyset$	
30	$R^0 \supset\sim (R^1 \equiv R^2)$	14	RU	{6}	M*
31	$\equiv R^0$	28,29,30	RU	{5,6}	M*
32	$\sim R^0\&R^0$	7,31	RU	{5,6}	M*
33	$[E(R^0 \equiv\sim (R^1 \equiv R^2))\& \sim (R^0 \equiv\sim (R^1 \equiv R^2))]\vee$ $[E(Q^0 \equiv\sim (Q^1 \equiv Q^2))\& \sim (Q^0 \equiv\sim (Q^1 \equiv Q^2))]$	32	RD	$\emptyset$	

Note that some of the lines are immediately marked. This means that the formulas they contain are not yet derivable. They may become derivable at a later stage of the proof, when some lines are unmarked. After line 33 is derived, line 27 is no longer minimal. As we only take minimal DAB-consequences into consideration for the determination of the formulas that are marked, this means that the marks of the lines whose fifth element is {4} (i.e. lines 12 and 22) must be removed. The proof can be continued as follows:

34	P^2	1,16	RU	{5}	M*
35	$(P^1\&P^2) \supset P^0$	12	RU	{4}	

36	$P^1\&P^2$	11,34	RU	{5} M*
37	P^0	35,36	RU	{4,5} M*
38	$\sim P^0\&P^0$	8,37	RU	{4,5} M*
39	$[E(P^0 \equiv (P^1\&P^2))\& \sim (P^0 \equiv (P^1\&P^2))]\vee [E(Q^0 \equiv\sim (Q^1 \equiv Q^2))\& \sim (Q^0 \equiv\sim (Q^1 \equiv Q^2))]$	38	RU	$\emptyset$

When line 39 has been derived, the set U_{39}*(Γ) (as defined in definition 3*) contains only the formula $Q^0 \equiv\sim (Q^1 \equiv Q^2)$. Since $E(Q^0 \equiv\sim (Q^1 \equiv Q^2))$ occurs at line 5, only the lines whose fifth element contains the number 5 must remain marked. The other marks (at lines 14, 19 and 30) must be removed. The diagnosis "b is not an XOR-gate", formally written as $\sim (Q^0 \equiv\sim (Q^1 \equiv Q^2))$, can now be conditionally derived in various ways:

40	$\sim (Q^0 \equiv\sim (Q^1 \equiv Q^2))$	12,39	RU {4}
40′	$\sim (Q^0 \equiv\sim (Q^1 \equiv Q^2))$	14,33	RU {6}
40″	$\sim (Q^0 \equiv\sim (Q^1 \equiv Q^2))$	12,14,27	RU {4,6}

Since no other minimal DAB-consequences are derivable, the marks are final.

The example shows that AL*EXP not only produces a diagnosis (the diagnosis is finally derivable) but also describes how we reason after a diagnosis has been made. Indeed, the lines that are never marked, and the lines that are marked at some stage but are nevertheless finally derivable, are conclusions we can still draw from the revised system description in which $E(Q^0 \equiv\sim (Q^1 \equiv Q^2))$ is replaced by $\sim (Q^0 \equiv\sim (Q^1 \equiv Q^2))$.

3.6 Our definition of non-explanatory diagnosis contains two conditions that select the most parsimonious options among the potential diagnoses. The definition of U*(Γ) incorporates these conditions. This is why AL^*_{EXP} produces diagnoses as defined in 2.3. It is obvious that AL_{EXP} does not produce diagnoses: in this logic, line 39 would lead to the addition of more marks, instead of to the removal of marks (all lines that contain 4 as condition must be marked again). However, AL_{EXP} can be used to construct logics that produces diagnoses in a different sense, i.e. defined by other principles than parsimony. For instance, the gates may be made by different manufacturers, one being more reliable than the other. In such case there should be a preference for gates made by the unreliable manufacturer, rather than a preference for parsimony. This different preference can be incorporated in a different logic by means of a definition analogous to definition 3*. So these logics would be built on AL_{EXP} in the same way as AL^*_{EXP} is built on this basic logic.

4. *Formal Analysis of Explanatory Diagnostic Reasoning in Systems*

In 4.1 we discuss the reasoning process that leads to weak explanatory diagnoses. In 4.2–4.4 we discuss the reasoning process that leads to strong explanatory diagnoses.

4.1 When weak explanatory diagnoses are sought for their own sake (i.e. not as a step towards a strong diagnosis), the reasoning process is abductive and fits the following scheme:

(AA) (1) We observe that Q and want an explanation for this phenomenon.
(2) We know that if P would be true, this would (together with background knowledge R) explain Q.
(3) We know that R is true.
(4) Because of (1)–(3), we decide to accept P as true.

"AA" stands for "abductive argumentation". An example of such reasoning would be to conclude that b is an IMPL-gate. The background knowledge is SD\Γ, i.e the original system description minus the claim that b is an XOR-gate. If b is an IMPL-gate, this explains (together with SD\Γ) our observations of the circuit.

The background knowledge is selected by means of the non-explanatory diagnosis: the background knowledge is always a contraction of the original system description, and the non-explanatory diagnosis determines which elements are removed from the SD. This entails that non-explanatory diagnoses are useful (even necessary) steps in the construction of a weak explanatory diagnosis.

4.2 Constructing a strong explanatory diagnosis is a three stage process: first we construct a weak explanatory diagnosis; then we ask questions and try to answer them; finally, we formulate a strong diagnosis based on the answers to the questions.

In this section we discuss the first stage. Like in 4.1, the reasoning process is abductive. However, we do not have to accept any statement as true. The abductive reasoning process therefore fits the following scheme:

(AHF) (1) We observe that Q and want an explanation for this phenomenon.
(2) We know that if P would be true, this would (together with background knowledge R) explain Q.

(3) We know that R is true.
(4) Because of (1)–(3) we decide to regard P as a hypothesis which deserves further investigation.

"AHF" stands for "abductive hypothesis formation".

4.3 The second stage in the construction of a strong diagnosis consists in formulating relevant questions (i.e. questions of which at least one possible answer eliminates at least one of the weak explanatory hypotheses) and attempts to answer these questions. First we have to group the explanatory hypotheses into an initial whether-question. Whether-questions are formally represented as ?{p_1, ..., p_n}, to be read as "Which of the statements p_1, ..., p_n is true?". Their presupposition is that p_1, ..., p_n are mutually exclusive and jointly exhaustive. In our example, the initial whether-question is:

(A) ?{IMPL(b), RIGHT(b), EQ(b), NOT-LEFT(b), AND(b), T_{14}(b), NEITHER(b), CONTR(b)}

An example of a relevant question would be:

(B) What if $\text{input}_1(b) = 1$ & $\text{input}_2(b) = 1$ & $\text{input}_1(a) = 1$?

In our example we have assumed that the answer is

$$\text{output}(c) = 1 \ \& \ \text{output}(a) = 0$$

This answer makes question (B) irrelevant, because we have an answer to it. It also makes question (A) irrelevant, because we now can ask a more specific question:

?{NOT-LEFT(b), T_{14}(b), NEITHER(b), CONTR(b)}

In order to formalise this question-answer process, we need a logic that is ampliative (because questions must be generated) and dynamic (because it must be possible to delete or mark questions when an answer is given or a more specific question can be asked).

Before we discuss the last stage, two general remarks must be made with respect to questions and answers:
(1) In many contexts, it will be impossible to answer all relevant questions. In such cases, a strong diagnosis is impossible.
(2) When we ask questions and perform measurements, it may happen that all the system descriptions we are considering are falsified. If our measurements lead to such situation, the non-explanatory diagnosis upon which our weak explanatory diagnoses were built, is mistaken. New weak diagnoses

can be generated from potential non-explanatory diagnoses that satisfy the conditions (c) and (d) of definition 2.3.

4.4 The non-explanatory diagnoses on which weak and strong explanatory diagnoses are built, can be mistaken. Therefore, the conclusion we draw after all but one of the weak explanatory hypotheses are eliminated by further experiments, does not follow deductively from our observations. As a consequence, the final argument by which the strong diagnosis is supported has the following format:

(IBE) (1) If P, then this explains (together with R) why Q is the case.
(2) $P \wedge R$ is better than any other explanation we have of Q.
(3) We observe that Q is the case.
(4) Because of (1)–(3) we accept that P is true.

"IBE" stands for "inference to the best explanation". There is no general criterion for what "better" is, but here it means "closer to the original system description". Like in 4.1, an ampliative dynamic logic is needed to formalise this reasoning process.

4.5 Unlike what we did in section 3, we will not attempt to develop a logic which can be used to formalise the three-stage reasoning process described in 4.2–4.4. This will be done in a separate paper.

5. *Diagnosis and diagnostic reasoning in individuals*

A well known and important example of diagnosis for faults in individuals can be found in the program INTERNIST-I, a program for medical diagnosis which makes use of techniques common in artificial intelligence. In section 5.1 we describe how this program works (cfr. Myers 1985 and Schaffner 1985b), while in 5.2–5.4 we use it to illustrate our general analysis of diagnoses for faults in individuals.

5.1 The program contains an extensive knowledge base for about 500 diseases. Every individual disease has its own disease profile, a list of manifestations that is associated with this specific disease. Let D be a disease and M a manifestation in the disease profile of D. The program's data base will then contain two clinical variables that link D and M. The first one is called *evoking strength*; this variable is given a number from 0 to 5 according to the answer given to the following question: "If the patient has M, how likely is it that he/she has D?" The interpretations of the values (cfr. Schaffner 1985b,

p. 15) are:

0 Nonspecific —manifestation occurs too commonly to be used to construct a differential diagnosis.
1 Diagnosis is a rare or unusual cause of listed manifestation.
2 Diagnosis causes a substantial minority of instances of listed manifestation.
3 Diagnosis is the most common but not the overwhelming cause of listed manifestation.
4 Diagnosis is the overwhelming cause of listed manifestation.
5 Listed manifestation is pathognomonic for the diagnosis.

Secondly, there is the *frequency* variable (with values from 1 to 5), which depends on the answer to the question: "If D is present, how likely is M?". The interpretations of the values are:

1 Listed manifestation occurs rarely in this disease.
2 Listed manifestation occurs in a substantial minority of cases of the disease.
3 Listed manifestation occurs in roughly half the cases.
4 Listed manifestation occurs in the substantial majority of the cases.
5 Listed manifestation occurs in essentially all cases (it is a prerequisite for the diagnosis).

Besides these two types of clinical variables, the knowledge base also contains a disease independent measure, valued from 1 to 5, for each manifestation. This measure is called the *import* and is decided on through the questions: "How important in general is this manifestation of disease in diagnosis? Must it be explained or can it be disregarded?".

Finally, the knowledge base also contains links between diseases: some diseases have a high probability of occurring together; some disease may presuppose another one.

When confronted with a diagnostic task, the program maps out a 'master list' of disease hypotheses, which all explain some or all of the given manifestations. To mark the match and lack of match between the patient and the theoretical disease profiles, each disease hypothesis gets an initial score. This score is calculated as follows:
(1) Credit is awarded according to the 'evoking strength' of the manifestations found in the patient which are explained by the disease hypothesis. A manifestation with evoking strength: 0 gives the disease 1 point, a manifestation with evoking strength 1 yields 4 points; similarly, 2 gives 10, 3 gives 20, 4 gives 40, and 5 gives 80.

(2) When certain manifestations are expected given the hypothesis, but found absent in the patient, a negative score is calculated according to the 'frequency' of these manifestations. The rules here are: $1 \Rightarrow -1$, $2 \Rightarrow -4$, $3 \Rightarrow -7$, $4 \Rightarrow -15$, $5 \Rightarrow -30$.
(3) Manifestations not accounted for, but found in the patient also count against the hypothesis. They are debited in accordance with the 'import' of the manifestations. Here the rules are $1 \Rightarrow -2$, $2 \Rightarrow -6$, $3 \Rightarrow -10$, $4 \Rightarrow -20$, $5 \Rightarrow -40$.
(4) A bonus is rewarded to each disease that is linked to a previously diagnosed disease.
Every hypothesis is marked with its proper score, by which it is ranked in a hierarchy of possible explanations for the given manifestations. The hypotheses above a certain threshold are kept in consideration.

If one of the hypotheses is in lead with a score-difference of 90 points or more, it is proposed as diagnosis. Otherwise a question-and-answer-session is started. When no competitors are in reach of 45 points of the best explanations, the *pursue-mode* is activated which asks questions with a high evoking strength for the topmost hypothesis. When five or more competitors are closer than 45 points, the *rule-out-mode* is activated which asks questions whose negative answers will result in the rapid elimination of some competitors. When less than five hypotheses stay in a distance of 45 points the *discriminative-mode* is activated, which asks questions to heighten quickly the score of one of the hypotheses and downgrade the scores of its competitors. In this mode both evoking strength and frequency numbers are taken in consideration.

The question and answer session usually results in a diagnosis (i.e. one of the diseases acquires a score that exceeds the scores of the competitors with at least 90 points). Two remarks are to be made in this respect:
(1) The program will of course also terminate when all questions are exhausted, even when multiple hypotheses are still to be taken into consideration.
(2) After proposing a diagnosis D, the program will look at the manifestations which are not explained by D. If all these manifestations have an import of two or less, the program stops. If some of the unexplained manifestations are more important, the program will assume that the patient has a second disease. To determine which one, only the manifestations unaccounted for by the first diagnosis are taken into account. In this way, multiple coexisting diseases can be accounted for.

5.2 Like in section 2, we need two preliminary definitions:

(5.1) T is a theory about individual i if and only if T is a set of non-quantified first-order sentences describing the *absence of* hidden properties in i.

(5.2) T∪OBS is an established fault of an individual characterised by theory T if and only if
(a) OBS is a set of first-order sentences describing observed properties (*manifestations*) of i, and
(b) T ∪ OBS is inconsistent.

In the case of medical diagnosis, the hidden properties are diseases. The theory T claims that all the diseases the program knows are absent. The observed properties are the manifestations. The manifestations contradict theory T, so a fault is established.

5.3 Non-explanatory diagnosis as defined in 2.3 presupposes that objects are analyzed into components. This is not done in theories as defined in 5.1, so this type of diagnosis is impossible here. What we do have is weak and strong explanatory diagnoses. Weak ones are defined as follows:

(5.3) Ω is a *weak explanatory* diagnosis for a fault in i if and only if
(a) it is of the form $\wedge \mathrm{P_j}i$ (where $\mathrm{P_j}$ is a predicate describing a hidden property),
(b) T contains the set $\Gamma = \{\neg \mathrm{P}_1 i, ..., \neg \mathrm{P_n} i\}$,
(c) $(\mathrm{T} \backslash \Gamma) \cup \Delta$ (deductively or inductively) explains (a part of) OBS, and
(d) every set Δ' satisfying the conditions (b) and (c) has at least as many elements as Δ.

The "master list" in the program contains such weak explanatory diagnoses: each disease in the list inductively explains a part of the manifestations, and the diagnoses are minimal (only one disease is postulated in the beginning).

Diagnoses as defined in 5.3 are generated by a process of abductive hypothesis formation which fits the following scheme:

(AHF*) (1) We observe that Q and want an explanation for this phenomenon.
(2) We know that if P would be true, this would (deductively or probabilistically) explain (a part of) Q.
(3) Because of (1) and (2) we decide to regard P as an hypothesis which deserves further investigation.

There are three differences between (AHF*) and (AHF) in section 4:

- In (AHF*) both deductive and inductive explanations are allowed, while in (AHF) explanations are assumed to be deductive.
- In (AHF*) partial explanations are allowed, while in (AHF) explanations are assumed to be complete.
- In (AHF*) there is no mentioning of background knowledge R.

The first two differences are accidental: it is possible to find examples of diagnostic reasoning in systems in which the explanations are probabilistic and partial (the definitions of section 2 can be adapted to include those cases). Conversely, partial and probabilistic explanations are characteristic of medical diagnosis but not of all cases of diagnosis for faults in individuals. The third difference is not accidental: background knowledge becomes necessary only if we analyze an object into components.

5.4 Strong explanatory diagnosis can be defined as follows:

(5.4) Ω is a *strong explanatory* diagnosis for a fault in i if and only if
(a) it is of the form $\wedge \mathrm{P_j} i$ (where $\mathrm{P_j}$ is a predicate describing a hidden property),
(b) T contains the set $\Gamma = \{\neg \mathrm{P_1} i, ..., \neg \mathrm{P_n} i\}$,
(c) $(\mathrm{T} \backslash \Gamma) \cup \Delta$ (deductively or inductively) explains (a part of) OBS, and
(d) every set Δ' satisfying the conditions (b) and (c) has at least as many elements as Δ.
(e) Δ scores much better that any other set satisfying the conditions (b)–(d).

Such strong explanatory diagnosis is obtained by (i) considering a set of weak diagnoses, (ii) formulating and answering relevant questions on the basis of this set, and (iii) drawing a final conclusion by means of inference to the best explanation. INTERNIST-I provides a clear illustration: the master list determines which questions are asked, and the proposed diagnosis is the best explanation; the operational criterion here is the 90 points difference.

Department of Philosophy and Moral Science
Ghent University
Blandijnberg 2
B-9000 Gent
Belgium
E-mail: Erik.Weber@rug.ac.be
Dagmar.Provijn@rug.ac.be

REFERENCES

Batens D. (2000), 'A Survey of Inconsistency-Adaptive Logics', in D. Batens, C. Mortensen, G. Priest & J.P. Van Bendegem (eds.), *Frontiers of Paraconsistent Logic*. London: Kings College Publications, pp. 49–73.

Batens D. (1998), 'Inconsistency-Adaptive Logics', in Eva Orlowska (ed.), *Logic at work. Essays Dedicated to the Memory of Helena Rasiowa.* Heidelberg & New York: Springer Verlag, pp. 445–472.

Myers J. (1985), 'The Process of Clinical Diagnosis and Its Adaptation to the Computer', in K. Schaffner (1985a), pp. 155–180.

Reiter R. (1980), 'A Logic for Default Reasoning', *Artificial Intelligence* 13, pp. 81–132.

Reiter R. (1987), 'A Theory of Diagnosis from First Principles', *Artificial Intelligence* 32, pp. 57–95.

Schaffner K. (ed.) (1985a), *Logic of Discovery and Diagnosis in Medicine.* Berkeley & Los Angeles, California: University of California Press.

Schaffner K. (1985b), 'Introduction', in K. Schaffner (1985a), pp. 1–32.

Weber E. & De Clercq K., (200+), 'Why the Logic of Explanation is Inconsistency-adaptive' in J. Meheus (ed.), *Inconsistency in Science.* Dordrecht: Kluwer Academic Publishers (in print).