Conflicts versus Analytical Redundancy Relations

A comparative analysis of the model based diagnosis approach from the Artificial Intelligence and Automatic Control perspectives

M-O. Cordier
IRISA, Campus de Beaulieu, F-35000 Rennes
cordier@irisa.fr
P. Dague, F. Lévy
LIPN-UMR 7030, Université Paris 13, 99 Avenue J-B. Clément, F-93430 Villetaneuse
{philippe.dague, francois.levy}@lipn.univ-paris13.fr
J. Montmain
EMA-CEA, Site EERIE - Parc George Besse, F-30035 Nîmes cedex1
jacky.montmain@site-eerie.ema.fr
M. Staroswiecki
LAIL-CNRS, EUDIL, Université Lille I, F-59655 Villeneuve d'Ascq cedex
marcel.staroswiecki@univ-lille1.fr
L. Travé-Massuyès
LAAS-CNRS, 7 Avenue du Colonel-Roche, F-31077 Toulouse cedex
louise@laas.fr

(as part of the French IMALAIA Group)

Submitted to the special issue of the IEEE SMC Transactions - Part B on

**Diagnosis of Complex Systems:
Bridging the methodologies of the FDI and DX Communities**

Abstract

Two distinct and parallel research communities have been working along the lines of the Model-Based Diagnosis approach: the FDI community and the DX community that have evolved in the fields of Automatic Control and Artificial Intelligence, respectively. This paper, which details and extends (Cordier *et al.*, 2000a, 2000b), clarifies and links the concepts and assumptions that underlie the FDI analytical redundancy approach and the DX logical approach. The formal match of the two approaches is proved and the theoretical proof of their equivalence together with the necessary and sufficient conditions is provided. This work results from the collaboration existing within the French IMALAIA group supported by the French National Programs on Automatic Control *GDR-Automatique* and on Artificial Intelligence *GDR-I3*, and *AFIA*.

I Introduction

Diagnosis is an increasingly active research topic, which can be approached from different perspectives according to the knowledge available. The so-called Model-Based Diagnosis (MBD) approach rests on the use of an explicit model of the system to be diagnosed. The occurrence of a fault is captured by discrepancies between the observed behavior and the behavior that is predicted by the model. Fault localization then rests on interlining the groups of components that are involved in each of the detected discrepancies. A definite advantage of this approach with respect to others, such as the relational approach (Peng, Reggia, 1990) or the pattern recognition approach (Dubuisson, 1990), is that it only requires knowledge about the normal operation of the system, following a consistency-based reasoning method.

Two distinct and parallel research communities have been using the MBD approach. The FDI community has evolved in the Automatic Control field from the seventies and uses techniques from control theory and statistical decision theory. It has now reached a mature state and a number of very good surveys exist in this field (Patton, Chen, 1991; Gertler, 1993; Frank, 1996; Iserman, 1997; CEP, 1997). The DX community emerged more recently, with foundations in the fields of Computer Science and Artificial Intelligence (Reiter, 1987; de Kleer, Williams, 1987, 1989; Hamscher, Console, de Kleer, 1992; Travé-Massuyès, Dague, Guerrin, 1997). Although the foundations are supported by the same principles, each community has developed its own concepts, tools and techniques, guided by their different modeling backgrounds. The modeling formalisms call indeed for very different technical fields; roughly speaking analytical models and linear algebra on the one hand and symbolic and qualitative models with logic on the other hand. The level of technicality of the contributions, and the fact that each community has its own set of conferences, publications and terminology, result in a poor understanding of the work in both sides.

The French IMALAIA group, supported by the French National Programs on Automatic Control *GDR-Automatique* and on Artificial Intelligence *GDR-I3* and AFIA, has been working along these lines, benefiting from the work already performed by the ALARM group (Cauvin *et al.,* 1998) and related work in France (Basseville, Cordier, 1996; Staroswiecki, 1998; Travé-Massuyès, Dague, 1999). The goals of this work are to agree upon a common DX/FDI terminology, to identify links in the concepts, similarities and complementarities in the DX and FDI methods, and to contribute to a unifying framework, thus taking advantage of the synergy of complementary techniques from the two communities.

This paper, which considerably details and extends (Cordier *et al*., 2000a, 2000b), clarifies and links the concepts that underlie the FDI analytical redundancy approach and the DX logical approach. In particular, the link between *structured parity equations or analytical redundancy relations* (ARR for short) and *conflicts* (in the sense of Reiter) is clarified by introducing the notions of *potential conflict* or *ARR support* and interpreting a conflict as the support of a non satisfied ARR. It is shown that the formal match of the two approaches can be proved provided completeness properties of the set of ARRs.

The FDI and DX approaches used for fault isolation are then analyzed from the two perspectives. It is shown that the first one, based on fault signatures, proceeds along a column interpretation of the fault signature matrix linking faults and ARRs whereas the later one, based on conflicts, proceeds along a row interpretation.

The results provided by the two approaches are then shown to be identical and the theoretical proof is included. This is proved in the no exoneration case under the single fault and the multiple fault assumptions, the exoneration case being left for further investigations. For the sake of clarity, the study is carried out in a pure consistency-based framework, i.e. without fault models.

The example that has been chosen to support the comparative analysis throughout the paper is the well-known system from (Davis, 1984) composed of three multipliers and two adders referred as the *polybox example* (figure 1). It refers to a typical static system, i.e. for which the transient behavior can be ignored. This choice has been made on purpose so that the comparison can be made in the classical framework of logical MBD and that the problems related to temporal diagnosis (Brusoni *et al.* 1998) can be ignored for now. Following the same idea, the only available observations are assumed to be those at one snapshot, putting aside the problems of incremental diagnosis and of the choice of the best next test point (de Kleer, Williams, 1987; de

Kleer *et al.,* 1991). In addition, the system is assumed to operate in an ideal non noisy and non disturbed environment.

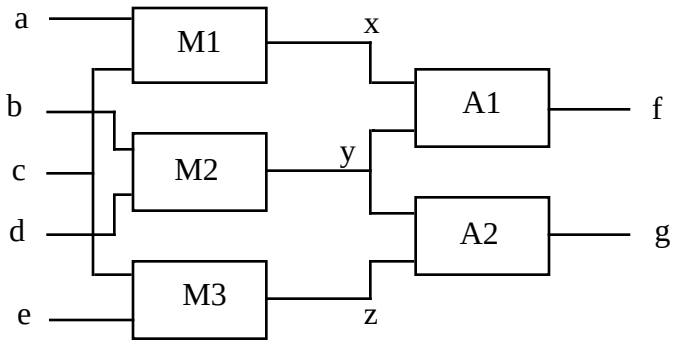

Figure 1 — *The system*

The paper is organized as follows. Section II presents the FDI analytical redundancy approach and the DX logical approach, respectively. Section III proposes a unified framework for the two approaches. The assumptions and concepts adopted by the FDI and DX communities are outlined and the correspondence between conflicts and ARRs is exhibited. Section IV proves the equivalence of the two approaches in the no exoneration case. Finally, Section V grounds the concepts into their implementation schemas and section VI discusses the results and outlines several interesting directions for future investigation.

II Presentation of the two approaches

II.1 The FDI analytical redundancy approach

II.1.1 The system model

A system is made of a set of components and a set of sensors, which provide a set of observations. The behavior model of the system expresses the constraints that link its descriptive variables. It is given by a set of relations, the formal expression of which depends on the type of knowledge (analytical, qualitative, production rules or numerical tables, etc.). It generally relies on a component-based description, which relates a set of constraints (or operators) to each component. In the case of the polybox example, elementary components are the adders A1, A2 (operators +), the multipliers M1, M2, M3 (operators ×) together with the set of sensors (identity operators).

Definition 2.1: The *system model* SM is defined as the behavioral model BM, i.e. the set of relations defining the system behavior, together with the observation model OM, i.e. the set of relations defining the observations that are performed on the system and the sensor models.

The set V of variables can be decomposed into the set of unknown variables X and the set of observed variables O. In the polybox example, we have:

$V = X \cup O$ where

$X = \{a, b, c, d, e, f, g, x, y, z\}$

$O = \{a_{obs}, b_{obs}, c_{obs}, d_{obs}, e_{obs}, f_{obs}, g_{obs}\}$

Behavioral Model (BM):

RM1: x = a × c
RM2: y = b × d
RM3: z = c × e
RA1: f = x + y
RA2: g = y + z

Observation model (OM):

RS*a*: $a = a_{obs}$
RS*b*: $b = b_{obs}$
RS*c*: $c = c_{obs}$
RS*d*: $d = d_{obs}$
RS*e*: $e = e_{obs}$
RS*f*: $f = f_{obs}$
RS*g*: $g = g_{obs}$

II.1.2 The diagnosis problem

The diagnosis requirements define a set of identifiers $\{F_{op}\}$ as the set of faults F that may occur on an operator op. Without loss of generality, we assume that there is a one-to-one correspondence between components and operators (see discussion in III.3) and the set of faults is hence noted $\{F_c\}$ where c is a component.

Definition 2.2: The set of observations OBS is a set of relations of the form v_{obs} = val, where $v_{obs} \in O$ and val is in the domain of v_{obs}.

In the polybox example, OBS = $\{a_{obs} = 2, b_{obs} = 2, c_{obs} = 3, d_{obs} = 3, e_{obs} = 2, f_{obs} = 10, g_{obs} = 12\}$ is a set of observations.

Definition 2.3: A diagnosis problem is defined by the system model SM, a set of observations OBS, and a set of faults F.

II.1.3 Analytical Redundancy Relations

Definition 2.4: An *analytical redundancy relation* (ARR) is a constraint deduced from the system model which contains only observed variables, and which can therefore be evaluated from any OBS. It is noted r = 0, where r is called the *residual* of the ARR.

ARRs are used to check the consistency of the observations with respect to the system model SM. The ARRs are satisfied if the observed system behavior satisfies the model constraints.
ARRs can be obtained from the system model by eliminating the unknown variables.

Definition 2.5: For a given OBS, the instantiation of the residual r is noted *val(r, OBS)*, abbreviated as *val(r)* when not ambiguous. *Val(r, OBS) = 0* thus means that the observations satisfy the ARR.

In the polybox example, three redundancy relations are ARR1, ARR2 and ARR3 (see section V.1 for more details on the way these ARRs are obtained from a structural analysis):

ARR1: $r_1 = 0$ where $r_1 \equiv f_{obs} - a_{obs}.c_{obs} - b_{obs}.d_{obs}$
ARR2: $r_2 = 0$ where $r_2 \equiv g_{obs} - b_{obs}.d_{obs} - c_{obs}.e_{obs}$
ARR3: $r_3 = 0$ where $r_3 \equiv f_{obs} - g_{obs} - a_{obs}.c_{obs} + c_{obs}.e_{obs}$

ARR1, ARR2 and ARR3 are obtained from the models of M1, M2, A1; M2, M3, A2; and M1, M3, A1, A2, respectively. If we assume that the sensors are not faulty, then the ARRs can be rewritten as:

ARR1: $f - (a.c + b.d) = 0$
ARR2: $g - (b.d + c.e) = 0$
ARR3: $f - g - a.c + c.e = 0$

Note that any of ARR1, ARR2 or ARR3 can be obtained from the two other ones.

II.1.4 Signature matrix

Besides analytical redundancy relations, a fundamental concept in the FDI approach is that of *fault signature*. The theoretical signature of a fault can be viewed as the expected trace of the fault on the different ARRs, given the system model.

Definition 2.6: Given a set ARR of ARR_i: $r_i = 0$, with Card(ARR) = n, the (theoretical) signature of a fault F_j is given by the binary vector $FS_j = [s_{1j}, s_{2j}, ..., s_{nj}]^T$ in which s_{ij} is given by the following application:

s: $ARR \times F \rightarrow \{0,1\}$

$(ARR_i, F_j) \rightarrow$ $s_{ij} = 1$ if the component affected by F_j is involved in ARR_i

$s_{ij} = 0$ otherwise

The interpretation of some s_{ij} being 0 is that the occurrence of the fault F_j *does* not affect ARR_i, meaning that $val(r_i) = 0$. On the other hand, the interpretation of some s_{ij} being equal to 1 is that the occurrence of the fault F_j *is expected to* affect ARR_i, meaning that $val(r_i)$ is now expected to be different from 0. This interpretation implicitly assumes that the occurrence of F_j is observable on the result of the ARR_i, or, equivalently, that if ARR_i is satisfied, then F_j is not present. As it will be stated later more formally, this is known as the *single fault exoneration (SF-exo) assumption*.

Definition 2.7: Given a set ARR of n ARRs, the signatures of a set of faults $F = \{F_1, F_2, ..., F_m\}$ all put together constitute the so-called *signature matrix* FS of dimensions n×m.

In the polybox example, the signature matrix for the set of single faults corresponding to components A1, A2, M1, M2 and M3, respectively, is given by:

	F_{A1}	F_{A2}	F_{M1}	F_{M2}	F_{M3}
ARR1	1	0	1	1	0
ARR2	0	1	0	1	1
ARR3	1	1	1	0	1

II.1.5 Multiple faults

The case of multiple faults can be dealt with by expanding the number of columns of the signature matrix, leading to a total number of 2^m-1 columns if all the possible multiple faults are considered. The theoretical signature of a multiple fault is generally obtained from the signatures of single faults as explained below. Consider that F_j is a multiple fault corresponding to the simultaneous occurrence of k single faults $F_1,..., F_k$, then the entries of the signature vector of F_j are given by:

$s_{ij} = 0$ if $s_{i1} = s_{i2} = \ldots = s_{ik} = 0$

$s_{ij} = 1$ otherwise, i.e. if $\exists\, l \in \{1,.., k\}$ such that $s_{il} = 1$

In the polybox example, the signature matrix above extended to double faults (all signatures of triple faults and above are identical to (1,1,1)) is given by:

	F_{A1}	F_{A2}	F_{M1}	F_{M2}	F_{M3}	F_{A1A2}	F_{A1M1}	F_{A1M2}	F_{A1M3}	F_{A2M1}	F_{A2M2}	F_{A2M3}	F_{M1M2}	F_{M1M3}	F_{M2M3}
ARR1	1	0	1	1	0	1	1	1	1	1	1	0	1	1	1
ARR2	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1
ARR3	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1

The interpretation of multiple fault signature entries is the same as for single faults. Given the way multiple fault signatures are derived from single fault signatures, this interpretation implies that the simultaneous occurrence of several faults is not expected to lead to situations in which the faults compensate, resulting in the non-observation of the multiple fault. As it will be stated later more formally, this is known as the *multiple fault exoneration (MF-exo) assumption*, which is a generalization of the exoneration assumption defined for single faults.

II.1.6 Diagnosis

The diagnosis sets in the FDI approach are given in terms of the faults accounted for in the signature matrix. The generation of diagnosis sets is based on a column interpretation of the signature matrix. The ARRs are instantiated with the observed values OBS and the associated residuals are determined, providing an *observed signature*, which can be compared with the fault theoretical signatures. This comparison is stated as a decision-making problem.

Definition 2.8: The signature of a given observation OBS is a binary vector $OS = [OS_1,...,OS_n]^T$ where $OS_i = 0$ if and only if $val(r_i, OBS) = 0$ and $OS_i = 1$ otherwise.

The first step is to decide whether a residual value is zero or not, in the presence of noises and disturbances. This problem has been thoroughly investigated within the FDI community. It is generally stated as a statistical decision-making problem, making use of the available noise and disturbance models (Basseville, Nikiforov, 1993).
The second step is to actually compare the observed signature with the fault signatures. A solution to this decision-problem is to define a *consistency criterion* as follows:

Definition 2.9: An observed signature $OS = [OS_1,...,OS_n]^T$ is consistent with a fault signature $FS_j = [s_{1j},...,s_{nj}]^T$ if and only if $OS_i = s_{ij}$ for all *i*.

The consistency criterion has clear semantics and is therefore appropriate for comparing the obtained diagnosis results with the ones obtained by the logical approach (cf. section 3). In practical situations, this definition is too demanding, hence the FDI community generally uses a weaker *similarity-based consistency criterion* (Cassar, Staroswiecki, 1994).

Definition 2.10: The *diagnosis sets* are given by the faults whose signatures are consistent with the observed signature.

In the polybox example, the following results about single faults are obtained from the signature matrix of II.1.4 for different observed signatures:

$(OS_1,OS_2,OS_3) = (0,0,0)$ ⇔ no fault

$(OS_1,OS_2,OS_3) = (0,1,1)$ ⇔ A2 or M3 faulty

$(OS_1,OS_2,OS_3) = (1,0,1)$ ⇔ A1 or M1 faulty

$(OS_1,OS_2,OS_3) = (1,1,0)$ ⇔ M2 faulty

$(OS_1,OS_2,OS_3) = (1,1,1)$ ⇔ no single fault

The results about multiple faults are obtained from the extended signature matrix of II.1.5. Among the four first cases above, the observed signatures (0,0,0) and (1,1,0) give the same results and we have the following changes:

$(OS_1,OS_2,OS_3) = (0,1,1)$ ⇔ A2 or M3 or (A2 and M3) faulty

$(OS_1,OS_2,OS_3) = (1,0,1)$ ⇔ A1 or M1 or (A1 and M1) faulty

So far, the new double faults are supersets of single fault candidates; hence they are not considered. Considering multiple faults does not bring thus more information for the four first observed signatures. This is not the case for the (1,1,1) signature where double faults appear:

$(OS_1,OS_2,OS_3) = (1,1,1)$ ⇔ 8 double faults (all except (A1 and M1) and (A2 and M3)) and all faults of size ≥ 3 as supersets.

Another interesting point to note is that, in the polybox example, the same results are obtained for the three first observed signatures when the procedure is applied on ARR1 and ARR2 only:

$(OS_1,OS_2) = (0,0)$ ⇔ no fault

$(OS_1,OS_2) = (0,1)$ ⇔ A2 or M3 faulty

$(OS_1,OS_2) = (1,0)$ ⇔ A1 or M1 faulty

In these examples, the use of ARR3, associated with r_3, does not provide any more localization power. This is obviously not the case for the two last observed signatures, for which r_3 is needed to disambiguate the signature (1,1). It can be noticed that ARR3 was obtained from the combination of ARR1 and ARR2. The contribution of this kind of additional redundancy relations and the existence of a minimal set of ARRs is discussed in VI.1.

It is worth mentioning that the FDI community has developed a big amount of work for obtaining so-called *structured residuals*, which are designed so that every residual is sensitive to a subset of faults (Gertler, 1993; Staroswiecki *et al.*, 1993). This provides a specific structure to the signature matrix. The localization power of a set of residuals can be derived from the properties of the signature matrix structure. Another approach is to design so-called *directional residuals*, which are designed so that the occurrence of a given fault gives a particular direction to the residual vector (observed signature). These methods make the choice of a set of ARRs whose signatures are more relevant than others.

II.2 The DX logical diagnosis approach

(Reiter, 1987) proposed a logical theory of diagnosis. This theory is often referred to as diagnosis from first principles; i.e. given a description of a system together with observations of the system's behavior which conflict with the way the system is meant to behave, the problem is to determine those components of the system which, when not assumed to be operating normally, restore the consistency with the observed behavior.

This approach, also referred to as the consistency-based approach, was later extended and formalized in (de Kleer, Mackworth, Reiter, 1992). In the following we refer to the basic definition of (Reiter, 1987) without considering posterior extensions and refinements.

II.2.1 The system model

The description of the behavior of the system is component-oriented and rests on first-order logic. The components are those elements subject to faults and that are part of the diagnosis of the system.

Definition 2.11: A system model is a pair (SD, COMPS) where:

1. SD, the *system description*, is a set of first order logic formulas with equality.
2. COMPS, the components of the system, is a finite set of constants.

The system description uses a distinguished predicate AB, interpreted to mean abnormal. $\neg AB(c)$ with c belonging to COMPS hence describes the case where the component c is behaving correctly.

Example (polybox continued):

$$COMPS = \{A1, A2, M1, M2, M3\}$$

$$\begin{aligned} SD = \{ & ADD(x) \wedge \neg AB(x) \Rightarrow Output(x) = Input1(x) + Input2(x), \\ & MULT(x) \wedge \neg AB(x) \Rightarrow Output(x) = Input1(x) \times Input2(x), \\ & ADD(A1),\ ADD(A2),\ MULT(M1),\ MULT(M2),\ MULT(M3), \\ & Output(M1) = Input1(A1),\ Output(M2) = Input2(A1), \\ & Output(M2) = Input1(A2),\ Output(M3) = Input2(A2), \\ & Input2(M1) = Input1(M3)\ \} \end{aligned}$$

Let us note one aspect which differs somewhat from the description of the system in the FDI approach: with the distinguished predicate AB it is possible to link explicitly a physical component with the formulas describing its behavior and to make explicit the fact that the formulas describe the normal behavior of the component.
Formulas describing the behavior of the components are generally expressed by constraints and need a constraint solver to be processed. In the absence of such a constraint solver, they can be preprocessed by hand, e.g., the two first constraints above can be rewritten as:

$$\begin{aligned} \{ & ADD(x) \wedge \neg AB(x) \Rightarrow Output(x) := Input1(x) + Input2(x), \\ & ADD(x) \wedge \neg AB(x) \Rightarrow Input1(x) := Output(x) - Input2(x), \\ & ADD(x) \wedge \neg AB(x) \Rightarrow Input2(x) := Output(x) - Input1(x), \end{aligned}$$

$MULT(x) \wedge \neg AB(x) \Rightarrow Output(x) := Input1(x) \times Input2(x)$,
$MULT(x) \wedge \neg AB(x) \wedge Input2(x) \neq 0 \Rightarrow Input1(x) := Output(x) / Input2(x)$,
$MULT(x) \wedge \neg AB(x) \wedge Input1(x) \neq 0 \Rightarrow Input2(x) := Output(x) / Input1(x)$ }

II.2.2 The diagnosis problem

A diagnosis problem results from the discrepancy between the normal behavior of a system as described by the system model and a set of observations.

Definition 2.12: A set of observations OBS is a set of first-order formulas.

Definition 2.13: A diagnosis problem is a triple (SD, COMPS, OBS) where (SD, COMPS) is a system model and OBS a set of observations.

Note that this definition matches Definition 2.3 provided that each fault F corresponding to a set Δ of components is described by:

$\bigwedge_{c \in \Delta \subseteq COMPS} AB(c)$.

Example (polybox continued): Suppose the polybox is given the inputs a = 2, b = 2, c = 3, d = 3, e = 2 and it outputs f = 10, g = 12 in response. The set of observations is represented by:
OBS = {Input1(M1) = 2, Input2(M1) = 3, Input1(M2) = 2, Input2(M2) = 3, Input2(M3) = 2, Output(A1) = 10, Output(A2) = 12}.

II.2.3 Diagnosis

A diagnosis is a conjecture that certain components of the system are behaving abnormally. This conjecture has to be consistent with what is known about the system and with the observations. Thus, a diagnosis is given by an assignment of a behavioral mode, AB or ¬AB, to each component of the system in a way consistent with the observations and the model.

Definition 2.14: A *diagnosis* for (SD, COMPS, OBS) is a set of components $\Delta \subseteq$ COMPS such that: $SD \cup OBS \cup \{AB(c) \mid c \in \Delta\} \cup \{\neg AB(c) \mid c \in COMPS - \Delta\}$ is consistent. A *minimal diagnosis* is a diagnosis Δ such that $\forall \Delta' \subset \Delta$, Δ' is not a diagnosis

Following the principle of parsimony, minimal diagnoses are often the preferred ones.

Proposition 2.1: If every occurrence in the clausal form of $SD \cup OBS$ of an AB-literal is positive, which is in particular the case in the absence of fault models and of exoneration models, the minimal diagnoses are sufficient to characterize all the diagnoses, i.e. the diagnoses are exactly the supersets of the minimal diagnoses.

The particular case mentioned in proposition 2.1 corresponds to SD limited to correct behavioral models expressed as necessary conditions (i.e. $\neg AB(x) \Rightarrow CM$ as in the example), that is to the absence of explicit fault models (i.e. of the form $AB(x) \Rightarrow FM$), which is the case studied in this paper, and to the absence of exoneration models (i.e. of the form $CM \Rightarrow \neg AB(x)$, which express sufficient conditions of correctness and can be generally seen as very weak, non predictive, fault models).

By virtue of proposition 2.1, we will limit ourselves most of the time in this paper to minimal diagnoses.

II.2.3.1 R-conflicts

A direct way of computing diagnoses based on definition 2.14 is a generate and test algorithm where subsets of components are selected, generating minimal ones first, and tested for consistency. The obvious problem is the inefficiency of this method. A method based upon the concept of conflict set has been proposed and is at the basis of most of implemented DX algorithms. This concept has been introduced by (Reiter, 1987) and will be designated by R-conflict in this paper.

Definition 2.15: An *R-conflict* for (SD, COMPS, OBS) is a set of components $C = \{c1, ..., ck\} \subseteq$ COMPS such that $SD \cup OBS \cup \{\neg AB(c) \mid c \in C\}$ is inconsistent, i.e.: $SD \cup OBS \models \bigvee_{c \in C}$

AB(c). A *minimal R-conflict* is an R-conflict, which does not strictly include (set inclusion) any R-conflict.

An R-conflict can be interpreted as follows: one at least of the components in the R-conflict is faulty in order to account for the observations; or equivalently it cannot be the case that all the components of the R-conflict behave normally. On the last expression of definition 2.15, it can be seen that an R-conflict identifies with a positive AB-clause which is an implicate of the system description and the observations.

Example (polybox continued): The polybox with the observations as seen above (f = 10, g = 12) has the following minimal R-conflicts: {A1, M1, M2} and {A1, A2, M1, M3} due to the abnormal value of 10 for f. In the case f = 10 and g = 10, the two minimal R-conflicts are: {A1, M1, M2} and {A2, M2, M3}. In the case f = 10 and g = 14, the three minimal R-conflicts are: {A2, M2, M3}, {A1, M1, M2}, and {A1, A2, M1, M3}.

II.2.3.2 Computing minimal diagnoses using R-conflicts

Using these minimal R-conflicts, it is possible to give a characterization of minimal diagnoses, which provides a basis for computing them. This characterization is based on the minimal hitting set definition which follows:

Definition 2.16: A *hitting set* for a collection C of sets is a set $H \subseteq \bigcup \{S / S \in C\}$ such that $H \cap S \neq \{\}$ for each $S \in C$. A hitting set intersects each set of the collection. A hitting set is minimal if and only if no proper subset of it is a hitting set for C. Obviously, in order to compute the minimal hitting sets of a collection C of sets, only those elements in C which are minimal have to be considered.

Proposition 2.2: Δ is a minimal diagnosis for (SD, COMPS, OBS) if and only if Δ is a minimal hitting set for the collection of (minimal) R-conflicts for (SD, COMPS, OBS).

Example (polybox continued):
With f = 12 and g = 12, the only minimal diagnosis is {}.

With f = 10 and g = 12 as above, there are four minimal diagnoses obtained by computing the minimal hitting sets for the collection of minimal R-conflicts {{A1, M1, M2}, {A1, A2, M1, M3}} which are: $\Delta 1$ = {A1}; $\Delta 2$ = {M1}; $\Delta 3$ = {A2, M2}; $\Delta 4$ = {M2, M3}.

With f = 10 and g = 10, there are five minimal diagnoses obtained by computing the minimal hitting sets for the collection of minimal R-conflicts {{A1, M1, M2}, {A2, M2, M3}}. They are:
$\Delta 1$ = {M2}; $\Delta 2$ = {A1, A2}; $\Delta 3$ = {A1, M3}; $\Delta 4$ = {A2, M1}; $\Delta 5$ = {M1, M3}.

With f = 10 and g = 14, there are eight minimal diagnoses obtained by computing the minimal hitting sets for the collection of minimal R-conflicts {{A1, M1, M2}, {A1, A2, M1, M3}, {A2, M2, M3}}. They are:
$\Delta 1$ = {A1, A2}; $\Delta 2$ = {A1, M2}; $\Delta 3$ = {A1, M3}; $\Delta 4$ = {A2, M1}; $\Delta 5$ = {A2, M2}, $\Delta 6$ = {M1, M2}; $\Delta 7$ = {M1, M3}; $\Delta 8$ = {M2, M3}.

A more general characterization of conflicts and diagnoses, available with exoneration models and with fault models, can be found in (de Kleer, Mackworth, Reiter, 1992), allowing to get conflicts and diagnoses from prime implicates and prime implicants of the logical theory and giving then a way of computing diagnoses using a theorem prover. Our aim in this paper being to compare the basis of the FDI and DX approach in the absence of fault models, we do not consider these extensions of the theory and limit ourselves to the above definitions.

III Unified framework for the DX and FDI approaches

This section first discusses the different ways DX and FDI formulate the diagnosis problem and links the different objects that underlie the concept of fault on each side. The notion *of potential conflict* or ARR *support* is introduced and the formal match of the two approaches is obtained, proving that a conflict can be interpreted as the support of a non satisfied ARR. The matrix framework is then proposed as suitable to strictly compare both approaches.

III.1 System model (SM) vs. system description (SD)

Both FDI and DX approaches are model-based.

In FDI, the system model SM is composed of the behavior model BM and the observation model OM of the non faulty system. Behavioral laws are described in BM as constraints between variables (in general a set of differential algebraic equations). Most works in the FDI community do not explicitly use the concept of component, and BM describes the system as a whole, using e.g. state space models. When component based models are used, topological knowledge is implicitly included as shared variables. The observation model describes which system variables are available from the sensors and the sensor models. In the simplest cases, the behavioral law of a non faulty sensor just equals some variable to the sensor output (an observed variable belonging to *O*): $a = a_{obs}$.

Very often, the observation model OM is not present in DX. The equality $a = a_{obs}$ for each variable in *O* is thus implicitly assumed, and sensor faults are dealt with by considering sensors as components. In DX, the system description SD includes explicit topological knowledge and behavioral models of components. The main difference with FDI is that the assumption of correct behavior of a component, which supports its model, is explicitly coded thanks to the AB predicate. So, if F is a formula[1] describing the correct behavior of a component c, SM just contains F (which implicitly means that the behavior of ¬AB(c) is given by F) whereas SD explicitly contains the formula: ¬AB(c) ⇒ F. To achieve a suitable comparison framework, further developments assume that the following property holds.

SRE Property (System Representation Equivalence): Let SM and SD respectively be a FDI and a DX model of the same system. The SRE property is true if each formula of SM representing (part of) a behavioral law of a component or sensor c appears in the right-hand side of an implication in SD, the left-hand side of which is ¬AB(c) and conversely, SM is then simply obtained from SD by substituting False to all occurrences of the AB predicate.

In the following, by virtue of the SRE property, SM and SD are equally used. The restriction of SM (SD) to the behavioral law(s) of a set of components C is denoted by SM(C) (SD(C)).

III.2 FDI observations versus DX observations

In DX, the set of observations expresses as a set of first-order formulas. It is hence possible to express disjunctions of observations, which provides a powerful language. However, very often, only conjunctions of atomic formulas are used. In FDI, the observations are always conjunctions of equalities assigning a real value and/or possibly an interval value to an observed variable. In the following, to favor the comparative analysis, we do assume that we have the same observation language In both FDI and DX approaches, OBS is identical and made up of relations $a_{obs} = v$, which assign a value v to an observed variable.

III.3 FDI faults vs. DX faults

DX adopts a component-centered modeling approach and defines a diagnosis as a set of (faulty) components. In FDI the concept of component is not in general the central one. Whereas DX abstracts the diagnosis process at the component level, FDI deepens the analysis down to variables and parameters. FDI faults hence rather correspond to the DX concept of *fault mode*. In general, several parameters can be associated with a given component, giving rise to different fault modes. The difference is that FDI faults are viewed as deviations with respect to the models of normal behavior whereas in DX's logical view the faulty behavior cannot be predicted from the normal model and the involved parameters. For deterministic models, two kinds of deviations are considered (Gertler, 1998):

- in the system parameters, which may take values different from the nominal ones. These are referred to as *multiplicative faults*[2].

[1] F can be assumed to be written in first-order predicate calculus, even if in practice a constraint logic programming framework is frequently used, the truth value of F being thus evaluated with respect to a given semantics of the constraints in a given domain.

[2] with reference to their influence on the state variable vector in a state space model.

- in known variables associated to the sensors and actuators. These are referred to as *additive faults*[2].

As a consequence, the columns of the signature matrix are generally associated with variables and parameters. The link between additive/multiplicative faults and components is hence easy to establish (Staroswiecki, 2001): sensor and actuator faults are generally modeled as additive faults whereas system component faults are modeled as multiplicative faults.

Note that, in FDI, system parameters may be physical parameters when the models are issued from physical first principles, or so called structural parameters when, typically, the model is the result of black-box identification. Structural parameters have no straightforward physical semantics. However, in some cases, it is possible to establish the (non necessarily one-to-one) correspondence with physical parameters (Isermann, 1989). In the two cases, the model developer must be able to make the link between parameters and physical components if the goal is fault isolation. On the other hand, linking variables to sensors and actuators is straightforward.

Conversely, the DX approach could easily account for FDI fault models by expressing the model at a finer granularity level. For instance, considering a single-input single-output (static) component c whose behavior depends on two parameters θ_1 and θ_2 , the standard DX model given by:

$$\text{COMPONENT}(x) \wedge \neg \text{AB}(x) \Rightarrow \text{Output}(x) = f(\text{Input}(x), \theta_1, \theta_2)$$
$$\text{COMPONENT}(c)$$

could be replaced by:

$$\text{COMPONENT}(x) \wedge \text{PARAMETER1}(y) \wedge \text{PARAMETER2}(z) \wedge \neg \text{AB}(y) \wedge \neg \text{AB}(z) \Rightarrow \text{Output}(x) = f(\text{Input}(x), y, z)$$

$$\text{PARAMETER1}(\theta_1), \text{PARAMETER2}(\theta_2), \text{COMPONENT}(c)$$

The component-based DX approach can hence be generalized by allowing the set COMPS to include not only components (including sensors and actuators), but also parameters. This framework is adopted in the following, COMPS standing for the set of *generalized components*, in one-to-one correspondence with FDI faults.

III.4 ARRs vs. R-conflicts

In the two approaches, diagnosis is triggered when discrepancies occur between the modeled (correct) behavior and the observations (OBS). As seen in section II.2, in DX, diagnoses are generated from the identification of R-conflicts, where an R-conflict is a set of components the correctness of which supports a discrepancy. In the ARR framework, discrepancies come from ARRs, which are not satisfied by OBS.

The fundamental correspondence between ARRs and R-conflicts is now established using the following definitions and property.

Definition 3.1: The *support* of an analytical redundancy relation ARR_i is the set of components (columns of the signature matrix) with a non zero element[3] in the row corresponding to this ARR_i.

Definition 3.2: The *scope* of a component c_j is the set of ARRs (rows of the signature matrix) with a non zero element in the column corresponding to c_j.

In II.1.3, ARRs have been defined with respect to a syntactic property (observed variables), and sets of ARRs are supposed to be (in some cases, proven to be) complete, in the sense that they are sensitive to relevant faults. Note that proving this property in the general case amounts to prove a general diagnosability property of faults. We will take it as an assumption, to be proven for particular systems under consideration, and moreover make a distinction between the standard view of completeness in FDI and a view taking ARR supports into account.

[3] It will be seen later that an extension can be done so that the elements of the FS matrix can take a value different from 1, when not equal to 0.

ARR-d-completeness Property: A set E of ARRs is said to be d-complete if:
E is finite;
for any OBS, if SM $\cup$ OBS $\models \perp$, then $\exists$ $ARR_i \in E$ such that $\{ARR_i\} \cup OBS \models \perp$.

ARR-i-completeness Property: A set E of ARRs is said to be i-complete if:
E is finite;
for any set C of components, $C \subseteq COMPS$, and for any OBS, if $SM(C) \cup OBS \models \perp$, then $\exists$ $ARR_i \in E$ such that the support of ARR_i is included in C and $\{ARR_i\} \cup OBS \models \perp$.

It will be clear from the comparison that d-completeness guarantees detectability, and i-completeness aims at isolation.

Proposition 3.1: Assuming the SRE property, let OBS be a set of observations for a system modeled by SM (or SD). 1) Given an analytical redundancy relation ARR_i violated by OBS, the support of ARR_i is an R-conflict; 2) If E is a d-complete set of ARRs, then if there exists an R-conflict for (SD, COMPS, OBS), there exists an analytical redundancy relation $ARR_i \in E$ violated by OBS; 3) If E is i-complete, then given an R-conflict C for (SD, COMPS, OBS), there exists an analytical redundancy relation $ARR_i \in E$ violated by OBS whose support is included in C.

Proof:
1) By hypothesis, $\{ARR_i\} \cup OBS \models \perp$; since, if C is the support of ARR_i, ARR_i is a consequence of $SM(C)$, it follows that $SM(C) \cup OBS \models \perp$, i.e. C is an R-conflict.

2) Suppose now that an R-conflict has been detected and that E is d-complete. Since an R-conflict exists, $SM \cup OBS \models \perp$, and d-completeness gives an $ARR_i \in E$ such that $\{ARR_i\} \cup OBS \models \perp$.

3) Last, let C be an R-conflict and suppose that E is i-complete. By definition of R-conflicts, one has $SM(C) \cup OBS \models \perp$, and i-completeness gives the result.

In consequence, the support of an ARR can be defined as a *potential R-conflict* (cf. the related concept of possible conflict in (Pulido, Alonso, 2002)).

Corollary 3.1: If both the SRE property holds and the ARR-i-completeness holds, the set of minimal R-conflicts for OBS and the set of minimal supports of ARRs (taken in any i-complete set of ARRs) violated by OBS are identical.

Given SM, COMPS, OBS, the equivalence between really computed minimal R-conflicts for that OBS on the one hand and minimal supports of those really computed ARRs which are falsified by OBS on the other hand, depends both on the existence of a complete problem solver for DX (computation of prime implicates) and of a computable i-complete set of ARRs. Proposition and corollary 3.1 state the conditions under which a formal equivalence holds. This is a key point of the comparison between the FDI and DX approaches. Notice that corollary 3.1 was stated in (Cordier *et al*., 2000b) as proposition 4.1, omitting the condition of i-completeness. This statement was thus exact only in the cases where an i-complete set of ARRs exists ((Pulido, Alonso, 2002) suggested rightly that some conditions were needed, but gave only a sufficient condition of effective computability without any characterization and did not point out any concept similar to i-completeness). This is the case for example for linear algebraic equations, but it has not be proved in general. The completeness properties will be discussed more deeply in VI.1.

Example (polybox continued):
The potential R-conflicts are: $C1 = \{A1, M1, M2\}$ (support of ARR1), $C2 = \{A2, M2, M3\}$ (support of ARR2) and $C3 = \{A1, A2, M1, M3\}$ (support of ARR3).
With f = 10 and g = 12, ARR1 and ARR3 are not satisfied, which gives rise to the minimal R-conflicts C1 and C3.
With f = 10 and g = 10, ARR1 and ARR2 are not satisfied, which gives rise to the minimal R-conflicts C1 and C2.
With f = 10 and g = 14, ARR1, ARR2 and ARR3 are not satisfied, which gives rise to the minimal R-conflicts C1, C2 and C3.

III.5 The matrix framework

The FDI approach uses the signature matrix crossing ARRs in rows and sets of components in columns. It was shown in II.1 that, given an observation OBS, diagnosis is achieved by identifying those columns, which are identical (or closest with respect to a distance function) to the observed signature.

In the DX approach, it has been seen in II.2 that (minimal) diagnoses are obtained as (minimal) hitting sets of the collection of (OBS-) R-conflicts. From III.4 above, under the assumption of i-completeness, such R-conflicts can be viewed as the supports of those ARRs which are not satisfied by OBS, i.e. looking at the corresponding set of rows I. A (minimal) hitting set of the collection of R-conflicts can thus be viewed as a (minimal) set J of singleton columns (i.e. columns corresponding to one single component) such that each of the rows of I intersects at least one column of J (i.e. has a non zero element in this column).

It is thus quite natural to adopt this matrix framework as a formal basis on which to compare the two approaches.

Let ARR = $\{ARR_i / i = 1…n\}$ be a set, assumed to be i-complete, of ARRs and COMPS = $\{c_j / j = 1…m\}$ be the set of components of the system. FS = $[s_{ij}]_{i = 1…n,\ j = 1…m}$ is the signature matrix. The j^{th} column of FS is the signature of a fault on c_j and is noted FS_j.

Definition 3.3: Any observation OBS splits the set ARR into two subsets. The subset of ARRs which are violated, i.e. $\{ARR_i \equiv (r_i = 0) / val(r_i, OBS) \neq 0\}$, is defined as R_{false}. The subset of ARRs which are satisfied, i.e. $\{ARR_i \equiv (r_i = 0) / val(r_i, OBS) = 0\}$, is defined as $R_{true.}$ $R_{true} = ARR \setminus R_{false}$.

OBS is thus described through its signature *OS*, which is the binary column vector defined by: for all i = 1…n, $OS_i = 1$ if $ARR_i \in R_{false}$ and $OS_i = 0$ if $ARR_i \in R_{true}$. Note that this is equivalent to: $OS_i = Fa_{OBS}(ARR_i)$, where Fa_{OBS} stands for "not satisfied" and denotes the *falsity* value of the relation ARR_i with respect to OBS.

The FDI theory compares the observed signature to the fault signatures whereas DX considers each line corresponding to an ARR in R_{false} separately, isolating R-conflicts before searching for a common explanation. In the following, these approaches are called *column view* and *line view* respectively.

III.6 Multiple faults

Notice that, in the matrix framework proposed in III.5, the DX approach deals with multiple faults by implicitly considering sets of singleton columns. By default, there is no limitation on the number of possible simultaneous faults: minimal diagnoses are built as minimal hitting sets of the collection of minimal R-conflicts and are not limited in size. Single and multiple faults are thus dealt with in exactly the same framework.

In the FDI approach, the signature matrix FS above, made up of singleton columns, is generally used in the case of single fault assumption. As seen in II.1.5, dealing with multiple faults requires adding new columns to FS, corresponding to the considered multiple faults (a maximum of $2^{|COMPS|} - |COMPS| - 1$ if all possible multiple faults are considered).

The following notation is used for columns representing non empty subsets of COMPS: for J = $\{j1,...,jk\} \subseteq \{1,...,m\}$, let us note C_J the subset $\{c_j / j \in J\}$[4], and s_{iJ} the matrix element of FS at line i and column FS_J (meaning the column added for C_J representing a multiple fault).

In order to compare FDI and DX approaches in the case of multiple faults, it is needed to specify how multiple fault signature columns are built from single fault signature columns in the FDI framework. Each of these new columns FS_J must be derived from the set of singleton columns FS_{j1}, ..., FS_{jk} by applying a given algorithm to extend deterministically the signature matrix. The

[4] Component C_j is here straightforwardly identified to $C_{\{j\}}$.

section IV shows that this algorithm should depend on the assumptions that are made about the combination of the effects of the single faults.

The requirements for this combination law to match the DX approach are quite clear: viewing a hitting columns set $\{FS_{j1}, ..., FS_{jk}\}$ of the rows set I as a new column FS_J corresponding to $C_J = \{c_{j1}, ..., c_{jk}\}$, it results from the hitting set definition that each row of I must intersect the column FS_J if and only if it intersects at least one of the FS_{j1} columns. The column FS_J must have thus a non zero element in a given row i of I if and only if at least one of the FS_{j1} columns has a non zero element in row i, i.e. FS_J has to verify for all rows i in I:

$s_{iJ} \neq 0$ if and only if $\exists l\ 1 \leq l \leq k\ s_{ijl} \neq 0$ (FI property)

As in the FDI perspective, the extended matrix is computed for any possible set I of rows, the FI property has to hold for each row i and extended column FS_J.

It happens that this is actually how the theoretical signatures of multiple faults are generally obtained from the signatures of single faults in the FDI approach (cf. II.1.5) and it simply expresses the intuitive idea that a multiple fault may affect an ARR if and only if at least one of the single faults it is made up of may affect this ARR. This means that the scope of a multiple fault is the union of the scopes of its single fault constituents.

IV Comparing DX and FDI approaches: assumptions and results

This section makes an intensive comparison of the DX and FDI approaches. It is shown that every approach adopts different diagnosis exoneration assumptions by default. Under the same assumptions, in particular with no exoneration at all, it is shown that the results provided by both approaches are identical and the theoretical proofs are included.

IV.1 Exoneration assumptions for the comparison

The originality and the power of both the FDI and DX approaches result from the fact that they are based only on the correct behavior of the components: no model of faulty behavior is needed. Nevertheless, different assumptions are adopted by default by each approach, leading to different computations of the diagnoses. These assumptions concern the manifestations of the faults through observations.

The DX approach makes absolutely no assumption about how a component may behave when it is faulty. This is because this approach is only based on a *reductio ad absurdum* principle: any discrepancy between the correct model and the observations necessarily implies that a component is faulty. This ensures the fundamental property of the DX approach, i.e. its logical soundness. In the matrix framework, this means that, for any given OBS, only those rows (ARRs) which are not satisfied by OBS are considered: for each one, its support constitutes the associated R-conflict. Possible diagnoses (sets of faulty components) are built from these R-conflicts. However, the DX approach allows one to state an explicit exoneration assumption at the level of every component: assume any component, the model of which is satisfied in a given context, correct in this context. Beyond the default assumption of DX (nothing assumed about faulty behavior), this exoneration assumption is equivalent to state that the occurrence of any fault always manifests in the sense that a faulty component does not behave according to its corresponding model. This hypothesis is commonly expressed explicitly in SD by modeling components with biconditionals (relating the explicit correctness assumption and the functioning law). Note that, as conditions of proposition 2.1 are no more satisfied in this case, only minimal diagnoses are still characterized in terms of R-conflicts, a superset of a diagnosis being not in general a diagnosis. We do refer to this assumption as to the component-based exoneration (COMP-exo) assumption.

Definition 4.1 (COMP-exo assumption): If the correct behavioral model of a component is satisfied in a given context (given observation OBS and assumption of correct behavior of some given components), then this component is assumed to be correct in this context.

Conversely, the FDI approach is based on a direct reasoning about the effects of a fault (column), viewed as a non satisfaction of the correct behavioral model of the corresponding component, on the ARRs (rows). In addition to the obvious fact that a fault cannot affect an ARR which it is not in its scope, which is the direct reasoning used in DX, the idea is that a fault necessarily manifests

itself by affecting the ARRs in its scope, causing them not to be satisfied by any given OBS. Hence, given OBS, not only, as in DX, is any component in the support of a non satisfied ARR a fault candidate, but also any component in the support of a satisfied ARR is implicitly exonerated (satisfied rows are thus also used in the reasoning). In fact this result is not sound but rests on an ARR-based exoneration (ARR-exo) assumption, which is implicitly made in the FDI approach and has to be considered explicitly in order to compare the FDI approach with the DX approach.

Definition 4.2 (ARR-exo assumption): A set of faulty components necessarily shows its faulty behavior, i.e. causes any ARR in its scope not to be satisfied by any given OBS. Or, equivalently, given OBS, each component of the support of a satisfied ARR is exonerated, i.e. is considered as functioning properly.

In the following, the comparison between DX and FDI approaches is made only in the case of no-exoneration at all, i.e. no COMP-exo in DX (which is the default case) and no ARR-exo in FDI (which is not the default case). The comparison of the FDI ARR-exo assumption and the DX COMP-exo assumption has been made, relying on the concept of alibi (Raiman, 1992), but is out of the scope of this paper and will be published apart.

IV.2 The no-exoneration case

In this subsection, under the SRE property, the no-exoneration case is now given a formal account in the matrix framework previously introduced, in order to specify formally which (sets of) components have to be considered as diagnoses in each case.

From the matrix viewpoint, the fact that ARR_i, if satisfied by OBS, exonerates c_j appears (cf. II.1.4) in FS as $s_{ij} = 1$. In order to release the default ARR-exo assumption in the FDI approach, it is necessary to express that a faulty component may or may not affect the ARRs in its scope. To make the difference with the previous case, the symbol X can be used instead of 1 for this purpose. We can now represent the fact that c_j belongs to the support of ARR_i but is not necessarily exonerated when ARRi is satisfied by OBS, by $s_{ij} = X$. The semantics of $s_{ij} = X$ is thus: a fault in c_j can explain why ARR_i is not satisfied by OBS, but ARR_i may happen to be satisfied by OBS even when c_j is faulty (to be compared with the semantics of $s_{ij} = 1$: a fault in c_j implies that ARR_i cannot be satisfied by any OBS).

The generalized use of an exoneration assumption for each component of the support of each ARR is called the *exoneration case* and corresponds to the assumption by default of the FDI approach (elements of FS take their values in {0, 1}). As said above, in the present comparison, we consider only the total lack of exoneration, called the *no-exoneration case* (elements of FS take their values in {0,X}). In this later case, definitions 3.1 and 3.2 translate to: the *support* of an ARR_i is the set $\{c_j / s_{ij} = X\}$; the *scope* of a component c_j is the set $\{ARR_i / s_{ij} = X\}$.

IV.2.1 The single fault no-exoneration case (SF-no-exo case)

The column associated with the faulty component must have X in non satisfied rows and 0 or X in satisfied rows. In this column view, the matching of the observed signature with a fault signature is thus based on the fact that an X in the fault signature is consistent with either a 0 or a 1 in the observed signature. So, it is just like using only non satisfied rows: the faulty component must have X in each such row.

So acceptable diagnoses are those $\{c_j\}$ verifying:

$R_{false} \subseteq Scope(c_j)$ (CV-SF-no-exo)[5]

In the line view, $\{c_j\}$ is an acceptable diagnosis if it hits all the supports of not satisfied ARRs, that is to say:

$\forall i\ (ARR_i \in R_{false} \Rightarrow c_j \in Support(ARR_i))$ (LV-SF-no-exo)

(LV-SF-no-exo) and (CV-SF-no-exo) are straightforwardly equivalent, because each one is equivalent to: $\forall i\ (Fa_{OBS}(ARR_i) = 1 \Rightarrow s_{ij} = X)$.

[5] For explicitness purpose, the formulas corresponding to the different cases are labeled as explained: C/LV: Column/Line View, S/MF: Single/Multiple Fault, (no-)exo: (no) ARR-based exoneration, FI: Fault Interaction.

We have thus the result:

Theorem 4.1: Under the assumption of i-completeness, FDI single fault diagnoses in the ARR-no-exoneration case are identical to DX single fault diagnoses.

Example (polybox continued) Releasing the exoneration assumption in the polybox example leads to the following single fault signature matrix:

	F_{A1}	F_{A2}	F_{M1}	F_{M2}	F_{M3}
ARR1	X	0	X	X	0
ARR2	0	X	0	X	X
ARR3	X	X	X	0	X

The following results are then obtained:

With outputs f = 10 and g = 12, i.e. observed signature (1,0,1), there are 2 single fault diagnoses {A1} and {M1}.
With outputs f = 10 and g = 10, i.e. observed signature (1,1,0), there is only one single fault diagnosis {M2}.
With outputs f = 10 and g = 14, i.e. observed signature (1,1,1), there is no single fault diagnosis.
With outputs f = 12 and g = 12, i.e. observed signature (0,0,0), there are 5 single fault diagnoses.

These results obtained by FDI are identical to those obtained by DX (cf. II.2.3.2).

Let us remark also that, except in the case of normal observation (null observed signature), these results are the same as under the default exo assumption (cf. II.1.6). This is because, as each one of the ARRs can be derived from the other two, the observed signatures (1,0,0), (0,1,0) and (0,0,1) are physically impossible. But this would not be the case in general. For instance, it is not the case here for the normal observation f=12, g=12, i.e. observed signature (0,0,0): in the exo case (cf. II.1.6), no single fault diagnosis exists, when in the no-exo case, five single-fault diagnoses corresponding to the five components are proposed.

IV.2.2 The multiple fault no-exoneration case (MF-no-exo case)

In this case, (CV-SF-no-exo) can be straightforwardly extended to: C_J is a possible diagnosis iff

$R_{false} \subseteq Scope(C_J)$ (CV-MF-no-exo)

No COMP-exo and multiple faults is the default case in DX. The way the line view selects a set of column vectors (cf III.6) to build the equivalent extended matrix column interprets as follows: a multiple fault can explain that a given ARR is not satisfied if and only if at least one of its faults can explain it, i.e. several faults never produce more than the combination of their separate effects. On the other hand, it is admitted that a faulty component does not necessarily affect an ARR in its scope (no-exo) and that several faults may compensate each other (compensation), resulting in a satisfied ARR.

With the help of the ordering 0<X, the no-exoneration fault interaction law can be stated very simply:

$s_{iJ} = \sup_{j \in J} \{s_{ij}\}$ (FI-MF-no-exo)

Thus in the line view the diagnoses are the sets C_J such that:

$\forall i\ (ARR_i \in R_{false} \Rightarrow \exists j \in J, C_j \in Support(ARR_i))$ (LV-MF-no-exo)

This, due to (FI-MF-no-exo), translates to:

$\forall i\ (ARR_i \in R_{false} \Rightarrow C_J \in Support(ARR_i))$

that in turn is the same as $R_{false} \subseteq Scope(C_J)$, i.e. (CV-MF-no-exo).

Theorem 4.2: Under the assumption of i-completeness, FDI diagnoses in the ARR no-exoneration case are identical to non empty DX diagnoses.

Example (polybox continued): For the polybox example, the following extended signature matrix is obtained from the usual one (see II.1.5) by replacing each 1 by X (all signatures of at least triple faults are identical to (X,X,X)):

	F_{A1}	F_{A2}	F_{M1}	F_{M2}	F_{M3}	F_{A1A2}	F_{A1M1}	F_{A1M2}	F_{A1M3}	F_{A2M1}	F_{A2M2}	F_{A2M3}	F_{M1M2}	F_{M1M3}	F_{M2M3}
ARR1	X	0	X	X	0	X	X	X	X	X	X	0	X	X	X
ARR2	0	X	0	X	X	X	0	X	X	X	X	X	X	X	X
ARR3	X	X	X	0	X	X	X	X	X	X	X	X	X	X	X

The following results are then obtained:

With outputs f = 10 and g = 12, i.e. observed signature (1,0,1), there are 4 minimal diagnoses: the 2 single fault diagnoses {A1} and {M1} and the 2 double fault diagnoses {A2, M2} and {M2, M3}, and 22 superset diagnoses.

With outputs f = 10 and g = 10, i.e. observed signature (1,1,0), there are 5 minimal diagnoses: the single fault diagnosis {M2} and the 4 double fault diagnoses {A1, A2}, {A1, M3}, {A2, M1} and {M1, M3}, and 20 superset diagnoses.

With outputs f = 10 and g = 14, i.e. observed signature (1,1,1), there are 8 minimal double fault diagnoses: {A1, A2}, {A1, M2}, {A1, M3}, {A2, M1}, {A2, M2}, {M1, M2}, {M1, M3} and {M2, M3}, and 16 superset diagnoses.

These results obtained by FDI are identical to those obtained by DX (cf. II.2.3.2). In the case where f = 12 and g = 12, i.e. observed signature (0,0,0), any non empty subset of components is a diagnosis: there are 5 minimal single fault diagnoses and 26 superset diagnoses. The only difference between FDI and DX is that, in this case, DX proposes also the empty diagnosis, {}, when FDI considers only the possible faults (it could be possible to add a "no-fault" column of signature (0,0,0) to the signature matrix in order to represent the empty diagnosis, and thus the results would be identical).

It can be noticed that, except in the f = 10 and g = 14 case (where anyhow, no exoneration can apply as no ARR is satisfied), these results are different from those obtained under the default exo assumption (II.1.6).

V Implementation issues

This section first presents the classical implementation schemes for the FDI and DX theories and shows the implications and how the two theories result in significant differences on the operational side.

V.1 Implementation of the FDI approach

This subsection presents the general lines of the so-called structural approach (Cassar, Staroswiecki, 1994) to obtain a set of ARRs and the signature matrix for a given diagnosis problem (SM, OBS, F), as defined in section II.1.2. Note that alternative methods have been proposed, based on different computational tools but providing equivalent outcomes (Carpentier 1999), (Frank, 1996), (Gertler, 1998), (Patton et al, 2000). Given a system modeled by SM, the system structure is defined through a binary application *st*:

$$st: R \times V \rightarrow \{0,1\}$$
$$(f, v) \rightarrow st(f, v)$$

where $st(f, v) = 1$ if and only if v appears in relation f, R and V being the set of relations in SM and the set of variables respectively.

Definition 5.1: The set of (f, v) pairs such that $st(f, v) = 1$ is called the *system structure*.

The system structure can be represented by a bi-partite graph $G = (R \cup V, A)$ in which the nodes are relation-nodes or variable-nodes and there exists a non-oriented edge a_{ij} between node f_i and

node v_j if and only if $st(f_i, v_j) = 1$. Equivalently, it can be represented by the incidence matrix of G. In the polybox example, this matrix is given by:

	a	b	c	d	e	f	g	x	y	z	a_{obs}	b_{obs}	c_{obs}	d_{obs}	e_{obs}	f_{obs}	g_{obs}
RM1	1		1					1									
RM2		1		1					1								
RM3			1		1					1							
RA1						1		1	1								
RA2							1		1	1							
RSa	1										1						
RSb		1										1					
RSc			1										1				
RSd				1										1			
RSe					1										1		
RSf						1										1	
RSg							1										1

The procedure can be guided by a structural analysis performed on the system structure, which aims at exhibiting the calculation paths of unknown variables from observed variables. This problem can be formalized in a graph theoretical framework, which comes back to the well-known problem of finding a complete matching[6] with respect to the unknown variables X in the bi-partite graph $G = (R \cup V, A)$. In the system structure matrix representation, a complete matching with respect to X (CM_X matching for short) appears as a selection of one and only one entry per column corresponding to an unknown variable, and per row, the rows corresponding to the SM relations.

The following is a CM_X matching on the system structure of the polybox example (selected entries are indicated by ⊕), after an appropriate permutation of rows and columns that exhibits over-constrained subsystems (i.e. subsystems for which the number of constraints is larger that the number of unknown variables they link). In this case, there is only one over-constrained subsystem, outlining two redundant relations RA1 and RA2. The existence of over-constrained subsystems is stated by the FDI community as a necessary condition for the system to be *monitorable* (Staroswiecki, Declerck, 1989).

	a	b	c	x	d	y	e	z	f	g
RSa	⊕									
RSb		⊕								
RSc			⊕							
RM1	1		1	⊕						
RSd					⊕					
RM2		1			1	⊕				
RSe							⊕			
RM3			1				1	⊕		
RSf									⊕	
RSg										⊕
RA1				1		1			1	
RA2						1		1		1

Once a CM_X matching has been found, every relation involved can be interpreted as a mechanism which determines the value of its matched variable from the other variables which appear in the relation[7]. The unknown variables are then wholly determined by the relations involved in the

[6] A matching in a bi-partite graph $G = (R \cup V, A)$ is a sub-graph $K = (R \cup V, A')$ such that the projections $P_R(A')$ and $P_V(A')$ of A' on R and V are one-to-one. It is complete with respect to a subset of variables X if and only if $X \subseteq P_V(A')$.

[7] Possible restrictions related to invertibility issues of the SM relation operators can be dealt with by marking the corresponding edges, which are then not allowed for selection or allowed under restrictive conditions.

matching. The remaining relations are hence redundant. ARRs are obtained from redundant relations by replacing the unknown variables by their symbolic expression in terms of observed variables, which can be guided by following the calculation paths indicated by the matching (this symbolic propagation is not generally made explicit by the FDI community, although ARR design methods like parity space or elimination theory based approaches perform it automatically). A CM_x matching indeed allows one to direct the edges of the bi-partite graph *G*. If the edges belonging to the matching are oriented from relations to variables and the remaining edges from variables to relations, then the resulting alternated oriented chains provide the unknown variables calculation paths.

In the polybox example, the previous CM_x matching results in two ARRs. Let us consider the first redundant equation RA1, which leads to the first ARR: ARR1.

RS*a*: $a = a_{obs}$
RS*c*: $c = c_{obs}$
RM1: $x = a \times c$
RS*b*: $b = b_{obs}$
RS*d*: $d = d_{obs}$
RM2: $y = b \times d$
RA1: $f = x + y$
RS*f*: $f = f_{obs}$

ARR1: $r_1 = 0$ where $r_1 \equiv f_{obs} - a_{obs}.c_{obs} - b_{obs}.d_{obs}$

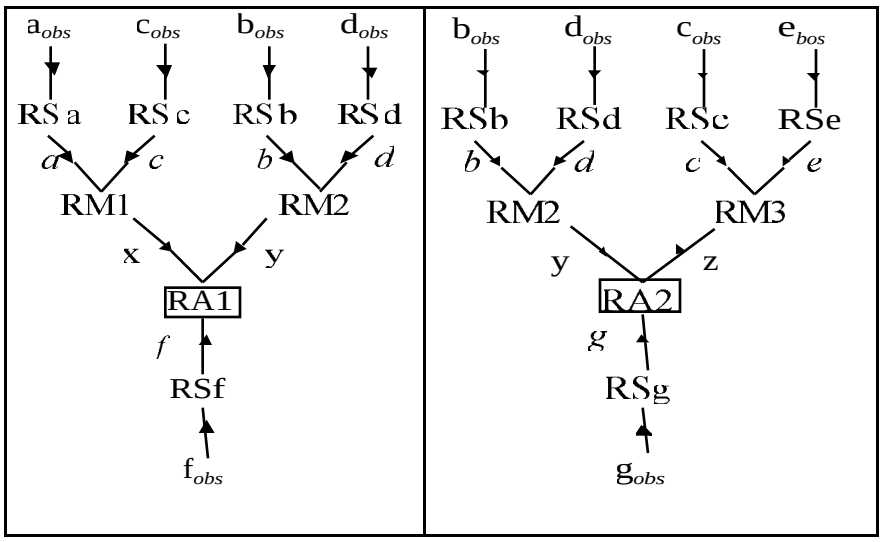

Figure 2 — *Graphical interpretations of ARR1 and ARR2*

Let us now consider the second redundant equation RA2, which leads to the second ARR: ARR2.

RS*b*: $b = b_{obs}$
RS*d*: $d = d_{obs}$
RM2: $y = b \times d$
RS*c*: $c = c_{obs}$
RS*e*: $e = e_{obs}$
RM3: $z = c \times e$
RA2: $g = y + z$
RS*g*: $g = g_{obs}$

ARR2: $r_2 = 0$ where $r_2 \equiv g_{obs} - b_{obs}.d_{obs} - c_{obs}.e_{obs}$

If we assume that the sensors are not faulty, then the two ARRs can be rewritten as:

ARR1: $f - (a \times c + b \times d) = 0$
ARR2: $g - (b \times d + c \times e) = 0$

Let us call the ARRs that are obtained from the perfect matching elementary ARRs. The number of elementary ARRs is the same for any existing perfect matching, as it only depends on the redundancy degree of the observed system (Luong et al., 1994). Given a set of elementary ARRs, additional redundancy relations can be obtained by combining the elementary ones. This is (algebraically) equivalent to substituting the expression derived from one relation for a variable in another relation. These additional redundancy relations correspond to those that would be obtained as elementary relations from other perfect matching existing on the system structure. In the polybox example, a third redundancy relation ARR3 can be obtained as follows:

RS*f*: $f = f_{obs}$
RS*a*: $a = a_{obs}$
RS*c*: $c = c_{obs}$
RM1: $x = a \times c$
RA1: $y = f - x$
RS*e*: $e = e_{obs}$
RS*c*: $c = c_{obs}$
RM3: $z = c \times e$
RA2: $g = y + z$
RS*g*: $g = g_{obs}$

ARR3*: r3 = 0 with* $r3 \equiv f_{obs} - g_{obs} - a_{obs}.c_{obs} + c_{obs}.e_{obs}$

The number of existing complete matching in a bi-partite graph is a source of combinatorial complexity. However, algorithms with only polynomial complexity have been proposed (Berge, 1975). It should be noticed that, given a set of ARRs, any combination of ARRs is also an ARR. This is the case in the previous example in which ARR3 is obtained from ARR1 and ARR2 and $r_3 = r_1 - r_2$. The number of ARRs may be relevant and this issue will be discussed in VI.1.

V.2 Implementation of the DX approach.

This section presents the general lines of the most common implementation of the DX approach. Most existing DX diagnosis systems are inspired from GDE (de Kleer, Williams, 1987) and use the ATMS (Assumption Based Truth Maintenance System) of de Kleer (1986) along the following architecture:

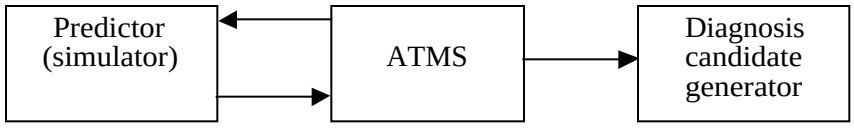

Figure 3 — *DX diagnosis systems classical architecture*.

The predictor is generally a constraint propagator. Constraints express the components behavioral laws. The basic inference step is to find a constraint that allows it to determine the value for a previously unknown variable. The newly recorded value may cause other constraints to trigger and more values to be deduced. A symptom is manifested when two different values are deduced for the same variable (i.e. logical inconsistency is detected). The dependencies tracing out the path through the constraints that the inputs (defined by OBS) have taken to deduce a given variable value prediction are used to construct the conflicts. Recording these dependencies is just the task of the coupled ATMS (cf. figure 3).

To every prediction is associated a *label* consisting of the disjunctive set of *minimal environments* from which it follows. An environment is a conjunctive set of *assumptions* (in our case these are correct behavior assumptions distinguished by the ¬AB symbol). A minimal inconsistent environment is referred to as a *nogood*. Hence a nogood exactly corresponds to an R-conflict. The ATMS records the inferences performed by the predictor as *justifications*, updates the labels of the facts according to these justifications and associated assumptions and keeps updated the set of nogoods.

Let us notice that the directionality of a component's signal flow is irrelevant to this constraint propagation based technique. Indeed, a component places a constraint between the values of its

terminals that can be used in any direction. For example, a substractor cannot be constructed by simply reversing an input and the output of an adder. But the constraint $f = x + y$ corresponding to A1 of our polybox example can be used in all directions, i.e. $x = f - y$ and $y = f - x$.

Consider the polybox example after the measurements $f = 10$ and $g = 12$. $y = 6$ can be calculated in two different ways: $y = b \times d = 6$, assuming that M2 is correct; $y = g - z = g - (c \times e) = 6$, assuming that both A2 and M3 are correct. The supporting environments of $y = 6$ are hence {M2} and {A2, M3}. On the other hand, we can calculate $y = f - x = f - (a \times c) = 4$, assuming that both A1 and M1 are correct. Hence, the supporting environment of $y = 4$ is {A1, M1}. Since $y = 6$ and $y = 4$ are inconsistent, we obtain two R-conflicts {M2, A1, M1} and {A2, M3, A1, M1}.

The diagnosis candidate generator can then compute candidate diagnoses from conflicts by applying a hitting set algorithm as proposed in (Reiter, 1987), and corrected in (Greiner *et al*., 1989).

VI Benefits and perspectives arising from the unified framework

VI.1 The SRE and ARR-completeness properties

The SRE property is required to perform a sound comparison. Indeed, it imposes that the models SM and SD are isomorphic both from a semantic and a syntactic point of view.

The d-completeness property is a standard requirement of FDI, since if a fault cannot be detected by any ARR, either it is out of interest, or the system is unsound and needs more sensors. On the theoretical level, this property can be stated as follows. Let M(**x**, **o**) be the equation set that represents SM, where **x** and **o** denote the vectors of variables contained in X and O respectively. A d-complete set of ARRs is a finite equation set E(**o**) such that (we abbreviate M(**x**, **o**) consistent and E(**o**) consistent by M(**x**, **o**) and E(**o**) respectively):

$$\forall \mathbf{o}\ (\exists \mathbf{x}\ M(\mathbf{x}, \mathbf{o}) \Leftrightarrow E(\mathbf{o})).$$

It is clear from this that d-completeness results from elimination theory that preserves equivalence (notice that in (Krysander, Nyberg, 2002) such an equivalence is included in the definition of an ARR, i.e. only d-complete ARRS are considered). An old result of algebraic geometry (Hodge, Pedoe, 1952) states that this property holds for *polynomial algebraic* equations. In this case, the theorem 4.2 applies and gives an equivalence between FDI (in the no-exo case) and DX diagnoses. But the result just ensures existence and is not constructive, and thus cannot be used in practice to build E. Progress has been made in this direction by computer algebra techniques (such that the use of Gröbner bases). In a general way, the d-completeness property is satisfied by sets E of ARRs which contain a basis (in the sense of a vector space basis) of all ARRs which can be built to describe a system. The theoretical conditions under which such a basis exists are related to the implicit functions theorem and can be exhibited in the form of a Jacobian condition in the case of *polynomial differential algebraic* equations (Staroswiecki, Comtet-Varga, 2001).

The i-completeness property is a novel concept since it requires to take into account the ARRs' supports. With the same notations as above, it can be stated as:

$$\forall M' \subseteq M\ \exists E' \subseteq E\ \forall \mathbf{o}\ (\exists \mathbf{x}\ M'(\mathbf{x}, \mathbf{o}) \Leftrightarrow E'(\mathbf{o})).$$

This is not common in the FDI community. The problem is related to the fact that having a basis of ARRs does not guarantee that all the potential minimal R-conflicts are represented by the ARRs' supports.

Example (polybox continued): The polybox example perfectly illustrates the above issue (cf. II.1.3). ARR3 may be obtained by combining ARR1 and ARR2, however its real support is {A1, A2, M1, M3} which does not include M2 whereas both Supp(ARR1) and Supp(ARR2) include M2.

The support may have to be taken into account relatively even to one single ARR, when nonlinear equations are dealt with. Let us consider the following system:

Example (the inverted polybox):

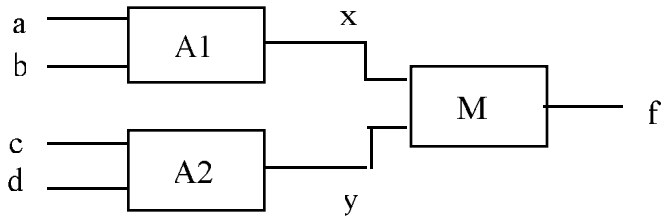

Figure 4 — *The inverted polybox*

Here COMPS = {A1, A2, M}, where A1 and A2 are adders and M is a multiplier, with models as in section II. We assume that $O = \{a, b, c, d, f\}$ and $X = \{x, y\}$.

The unique ARR is given by:

ARR1: $f - (a + b) \times (c + d) = 0$, with support {A1, A2, M}.

Let us consider the following observations: OBS = $\{a = -1, b = c = d = f = 1\}$. ARR1 is violated by OBS, giving rise to FDI single fault diagnoses {A1}, {A2} and {M}. But, if we consider the DX approach, then an R-conflict for OBS is {A1,M} because

SD $\cup$ {¬AB(A1)} $\cup$ $\{a = -1, b = 1\}$ |= $x = 0$
SD $\cup$ {¬AB(M)} $\cup$ $\{x = 0\}$ |= $f = 0$

and this conflict does not appeal to the behavior of A2. Thus DX single fault diagnoses are {A1} and {M} different from FDI ones.

It seems thus from theorem 4.1 that, even for such a simple system, ARR-i-completeness is not satisfied. Indeed, on the one hand SM({A1, M1}) $\cup$ OBS |= $\perp$, on the other hand the unique ARR actually satisfies {ARR1} $\cup$ OBS |= $\perp$, but Support(ARR1) = {A1, A2, M} is not included in {A1, M}.

The problem actually comes from a lack of precision in the definition of an ARR. Definition 2.4 just requires that an ARR is a constraint between variables belonging to *O* entailed by SM. But it does not precise what is the syntactic language of an ARR.

The spirit of FDI is that an ARR is a symbolic expression in terms of variables of *O*, obtained by symbolically eliminating the non observable variables of *X* between model equations. In particular no potential value of any observable should be used in an ARR, as it is computed before knowing any OBS. The problem arises when particular values of (observable or non observable) variables, when input in a component's model SM(C), determine its output independently of a subset of remaining inputs.

Thus, if the model of M is expressed as: $f = x \times y$ (1), it is quite natural to consider ARR1 as the only ARR. But the two particular cases that may yield a smaller support can be stated apart by augmenting the model of M of two (redundant) equations:

if $x = 0$ *then* $f = 0$ (2)
if $y = 0$ *then* $f = 0$ (3)

where the particular constant value 0 occurs (equation (2) is precisely the one that has been used above in DX propagation of OBS). To obtain i-completeness, it is natural to consider also as ARRs, in addition of ARR1:

ARR2: *if* $a + b = 0$ *then* $f = 0$, with support {A1, M}
ARR3: *if* $c + d = 0$ *then* $f = 0$, with support {A2, M}

obtained respectively by eliminating x between the model of A1 and equation (2) and by eliminating y between the model of A2 and equation (3). We call ARR2 and ARR3 (partially)

instantiated ARRs. Thus, ARR2 is violated by OBS and gives rise to FDI single faults {A1} and {M} as DX.

The key point is that, for the particular value 0 at one input, the model of the multiplier is expressed as a relation between two variables instead of three (the other input disappears). This leads to a strictly smaller support for the ARR, and thus to a strictly smaller potential conflict.

In the general case, let us consider a given ARR_i depending on a set of observed variables $\{o_1,...o_p\}$, then the partial derivative of ARR_i with respect to either of these variables is generically non zero. However since in non-linear systems, the partial derivatives are functions of the operating point, it can be understood that they may be zero or not, depending on the system's trajectories. Therefore the structure of the ARRs may change for those operating points for which some of the partial derivatives cancel, giving rise to as many instantiated ARRs.

The ARR-i-completeness issue is naturally linked to the redundancy and minimality issues. It is known in DX that only minimal (for subset inclusion) R-conflicts are relevant, the non minimal ones being redundant. On the other hand, we showed in V.1 that a given CM_x matching in the bi-partite graph of the system structure provides a set of elementary ARRs, but that other ARRs can be computed by combining these elementary ones or by using other existing CM_x matching. By construction, combined ARRs are clearly redundant when considered just as equations. But, in the light of the logical framework, one has to consider each ARR jointly with its associated support to obtain i-completeness. The question that must be answered is hence: under which conditions is a given ARR_j a logical consequence of a minimal set of ARR_is, $i \neq j$, and can thus be considered as redundant?

Proposition 6.1: The necessary and sufficient condition for a given ARR_j to be a logical consequence of a set of ARR_is, $i \in I$, $j \notin I$, is: $\exists\ I' \subseteq I$ such that
1) for any observation OBS, if all ARR_is, $i \in I'$, are satisfied by OBS, then ARR_j is satisfied by OBS (or, equivalently, if ARR_j is not satisfied by OBS, necessarily at least one of the ARR_is is not satisfied by OBS): $\bigwedge_{i \in I'} ARR_i[OBS] \Rightarrow ARR_j[OBS]$ is valid.
2) the support of ARR_j contains the support of each ARR_i, $i \in I'$:
$Supp(ARR_j) \supseteq \bigcup_{i \in I'} Supp(ARR_i)$.

ARR[OBS] designates the ground formula obtained from ARR by substituting each observed variable by its value in OBS: if OBS = $\{X_j = v_j\}$ then ARR[OBS] = ARR[X_j/v_j] (and so, what we designated by Fa_{OBS}(ARR) is nothing else that the falsity value of ARR[OBS], with respect to the semantics of the constraints of the domain).

Proof: These conditions are obviously sufficient: they traduce exactly that, each time ARR_j produces an R-conflict (i.e. is not satisfied by one given OBS), then one of the ARR_is produces a smaller or equal R-conflict. The proof that these conditions are necessary can be stated as follows. The logical form of a given ARR_j, making explicit the assumptions about the normality of components belonging to the support of ARR_j is the valid formula: $\bigwedge_{Ci \in Supp(ARR_j)} \neg AB(Ci) \Rightarrow ARR_j$, where free (observed) variables in ARR_j are assumed to be universally quantified (with respect to their interpretation domain). This can be equivalently formulated as the validity of: $PC(ARR_j) \vee ARR_j$, where $PC(ARR_j) = \bigvee_{Ci \in Supp(ARR_j)} AB(Ci)$ is the positive AB-clause formulation of the potential R-conflict $Supp(ARR_j)$. Suppose now that a given ARR_j is a logical consequence of a set of ARR_is, $i \in I$. This is logically expressed as:

$$\bigwedge_{i \in I} (PC(ARR_i) \vee ARR_i) \models (PC(ARR_j) \vee ARR_j) \qquad (5.1)$$

i.e.: $\forall\ L \subseteq I\ (\bigwedge_{l \in L} PC(ARR_l) \wedge \bigwedge_{k \in I \setminus L} ARR_k) \models (PC(ARR_j) \vee ARR_j)$.

Taking into account that $PC(ARR_l)$ and $PC(ARR_j)$ are AB-clauses and that ARR_k and ARR_j do not contain any AB-literal, this is equivalent to:

$$\forall L \subseteq I \ (\bigwedge_{l \in L} PC(ARR_l) \models PC(ARR_j)) \text{ or } (\bigwedge_{k \in I \setminus L} ARR_k \models ARR_j) \qquad (5.2)$$

Let $I' = \{i \in I \mid Supp(ARR_i) \subseteq Supp(ARR_j)\}$. For each element $i \in I \setminus I'$, there exists a component in $Supp(ARR_i)$ which does not belong to $Supp(ARR_j)$, i.e. there exists a positive AB-literal in the clause $PC(ARR_i)$ which does not appear in the clause $PC(ARR_j)$. Thus:

$$\bigwedge_{i \in I \setminus I'} PC(ARR_i) \not\models PC(ARR_j).$$

From this, one can deduce by choosing $L = I \setminus I'$ in (5.2) that:

$$\bigwedge_{i \in I'} ARR_i \models ARR_j$$

which is nothing else than condition 1 for $\{ARR_i\}_{i \in I'}$ and ARR_j.

As for any $i \in I'$ $Supp(ARR_i) \subseteq Supp(ARR_j)$, the condition 2 for $\{ARR_i\}_{i \in I'}$ and ARR_j is also satisfied. This ends the proof that the conditions 1 and 2 are necessary.

In FDI, any ARR which is obtained as a combination of elementary ARRis, say $\otimes_{i \in I} ARR_i$, practically satisfies condition 1 because any combination operator $\otimes$ is obviously required to verify $\bigwedge_i ARR_i[OBS] \Rightarrow \otimes_i ARR_i[OBS]$ for any OBS. However, computing trivially the union of the supports of the ARR_is does not give in general the right (i.e. minimal) support of $\otimes_i ARR_i$. Indeed, $Supp(\otimes_i ARR_i)$ is more often than not strictly included in $\cup_{i \in I} Supp(ARR_i)$, which makes that condition 2 is not satisfied in general and consequently $\otimes_i ARR_i$ is not redundant. The reason is that, in a combination $\otimes_i ARR_i$, some variables/parameters associated to some components are eliminated and it thus may happen that some components, which were involved in the supports of some ARR_i, are actually not needed to support $\otimes_i ARR_i$. In principle, the computation of the real support could be automated, thanks to the correspondence between variables/parameters and components, from symbolic computation of combination of elementary ARRs. In practice, both computations of the combinations and of their supports are either done by hand or automatically derived from other CM_x matching.

The existence of instantiated ARRs is, to the best of our knowledge, a novel issue, that we could outline from the comparison of the precompiled supports of ARRs and the on-line computed R-conflicts. As a matter of fact, instantiated ARRs cannot be derived from a pure structural analysis as the ones proposed in (Krysander, Nyberg, 2002) or (Pulido, Alonso, 2002). One could argue that such instantiated ARRs, which have to be considered when diagnosing discrete systems (e.g. if models are written in extension, all ARRs are instantiated), have few practical interest for continuous systems, those usually considered by FDI, because they occur only at singular points. This has to be further investigated on real examples. But in any case, instantiated ARRs are most probably relevant for hybrid systems. Establishing conditions under which instantiated ARRs exist and provide additional diagnosis information is thus a future direction of research, as is the possibility of obtaining i-completeness (at least, the set of possible supports for ARRs is finite, which opens to enumerative methods).

VI.2 Off-line vs. on-line computation of R-conflicts

From the computational point of view, the main difference between the FDI and DX approaches is that in FDI most of the computational work is done off-line. Using just the knowledge of which variables are observed, i.e. sensor locations, modeling knowledge is compiled: ARRs are obtained by combining model equations or constraints, and eliminating unobserved variables. The only thing that has to be done on-line, i.e. when a given OBS is acquired, is to compute the truth value (with respect to OBS) of each ARR and to compare the obtained observed signature with the fault theoretical signatures (columns of the signature matrix). In terms of R-conflicts, this means that

potential R-conflicts are compiled and that, for a given OBS, R-conflicts are exactly those potential R-conflicts which are supports of those ARRs which are not satisfied by OBS.

From a practical point of view, it is important to notice that computing ARRs is achieved by computer algebra techniques aiming at formally eliminating unobserved variables in equations. In FDI, this symbolic propagation process is not systematically automated but it is guided by a prior structural analysis (cf. V.1). The absence in SM of explicit assumptions (of components correctness) governing the validity of component models also implies that the support of each ARR, and thus the signature matrix, is computed by hand. Conversely, in DX, instantiated propagation is usually performed, i.e. the computational task starts as soon as OBS is known, nothing being compiled off-line. But the presence in SD of explicit assumptions supporting the validity of the models allows R-conflicts to be automatically derived by tracing the propagation process (usually using an ATMS coupled with the inference engine, cf. V.2).

It is important to notice that nothing in the logical formalism of DX prevents to use symbolic computation. In fact, the idea, coming from FDI, of compiling ARRs can be used as so in the DX framework for obtaining potential R-conflicts. In a second step, a given OBS is used to derive the R-conflicts in order to generate, as usual, diagnoses. Notice that this second step makes a quite different use of ARRs and OBS than the one adopted in FDI whereas the first step (symbolic compilation) is a matter of problem solvers. The FDI approach is generally limited to partial symbolic propagation. In order to deduce more potential R-conflicts, symbolic constraint solving methods would be needed. In some cases (e.g. linear equations), such complete algorithms, producing all possible ARRs and associated potential R-conflicts, can be implemented.

This has indeed already been proposed in the DX framework: in his thesis (Loiez, 1997) computes in advance all possible linear combinations of models eliminating all occurrences of unobserved variables, i.e. all possible ARRs, for the monostable circuit, an analog electronic circuit proposed as a reference problem (Dague, Taillibert, 1996). When this is possible, this is a way to get the best from each approach:

- modeling knowledge is compiled (under ARRs form) according to sensor locations before any observation has been made, which is the main advantage of the FDI approach;
- thanks to explicit correctness assumptions, potential R-conflicts (supports of ARRs) are computed at the same time to give rise, given an OBS, to R-conflicts;
- R-conflicts are used to generate the diagnoses, which is the main advantage of the DX approach.

Obviously, this requires that all observations are acquired at one go (or equivalently, that diagnosis begins when all observations have been made): there is no analog in FDI of the incremental evolution of diagnoses with respect to new observations and of the choice of the next observation in order to discriminate at best between present diagnoses.

But it is easy to imagine how to transpose these DX techniques, while keeping the benefit of the compilation. Given the location of all sensors, i.e. the set O of all observable variables, one can successively, for each subset O' of O, compile the set of ARRs whose variables are exactly the elements of O' and index it by O' and order these sets of ARRs according to the order of their indexes for set inclusion in the lattice of subsets of O. For a partial set of observations OBS' concerning a subset O' of observed variables, only those ARRs whose indexes are subsets of O' are evaluated, checked for consistency and possibly considered to generate R-conflicts. In a perspective of sequential observation, other ARRs, whose indexes intersect O', can be partially evaluated in advance and their indexes updated (elements of O' are removed). When adding a new observation concerning a variable o of $O \setminus O'$, only those ARRs whose updated indexes are $\{o\}$ have to be checked. Strategies to choose the best next observation, i.e. the best o in $O \setminus O'$ to be observed, like those commonly used in DX, can be applied. For example a one-step look-ahead strategy (de Kleer et al., 1991) just involves to check all ARRs whose updated indexes are singletons (these ARRs are constraints involving only one variable). Again, no variable propagation is needed at all.

Notice finally that, in the perspective of generating diagnoses from R-conflicts, what is compiled as supports of ARRs are the potential R-conflicts, not the minimal R-conflicts for a given OBS. In fact, as the potential R-conflicts are compiled before any observation has been made, it is impossible, once an observation is given, to ensure that only minimal R-conflicts for this observation are obtained. Even if an irredundant, i.e. logically minimal (no element is a logical

consequence of the others in the sense of proposition 6.1), and i-complete set of ARRs is available, it gives rise, for any given OBS, to a set of R-conflicts which are not generally all minimal. This is because a strict inclusion between supports of some of these ARRs is possible: it may happen that $Supp(ARR_i) \subset Supp(ARR_j)$ and that ARR_i and ARR_j are both not satisfied by OBS (remark that this does not contradict the set of ARRs being irredundant, which just implies the existence of at least another observation OBS' for which ARR_i is satisfied and ARR_j is not satisfied). However, if wanted, the generation of only minimal R-conflicts for any given OBS can be achieved. This just requires to compile the lattice (for set inclusion) of all potential R-conflicts, labeling each one by the set of ARRs of which it is the support. For a given OBS, it is then enough to explore this lattice from bottom in a breadth first manner, stopping to explore supersets each time a non satisfied ARR is met in the label of a potential R-conflict: one obtains this way exactly all minimal R-conflicts for OBS. This search could even be itself compiled, giving as result a direct mapping between all possible fault signatures and the associated sets of minimal R-conflicts. Given any OBS, only the observed signature has to be computed and minimal R-conflicts are thus directly obtained without any computation.

VI.3 Uncertainty management

It has been mentioned in II.1.6 that, when the observed signature fits no fault signature, some FDI applications accept the closest fault signatures using a similarity-based consistency criterion, e.g. with respect to some distance. The reason for accepting an approximate matching is that it is a way to cope with model uncertainties, e.g. unknown disturbances or model errors. As uncertainty is not fully characterized, the semantics of the distance is not clearly defined. However, the goal is to guarantee some kind of robustness with respect to the decision procedure, which assesses whether a residual is zero or not. For example, using the Hamming distance[8] as criterion, the correct fit of the observed signature is guaranteed in presence of k decision errors if the fault signatures are all at least at a distance of 2k+1 one from the other. Nevertheless, it seems that this operation can be viewed as hypothesizing a whole set of possible observed signatures and needs no change in the framework relating observed and fault signatures.

Another way to deal with uncertainties in FDI is to make use of as many ARRs as can be derived, even though these may be redundant from a detection and localization point of view. However, it can be argued that additional signature bits ensure more robust detection in the presence of noise and disturbances (like error code bits in information theory), and this suggests to proceed with all the available ARRs. Although a definition of *logically redundant ARRs* is provided in section VI.1, the redundancy properties of ARRs in noisy environments must be stated in statistical terms and are not studied in this paper.

The above mentioned methods used in FDI are among a set of methods which aim at providing the diagnosis system with robustness. This is indeed an issue arising from the type of models being used, which are essentially numeric with uncertainties represented either by unknown disturbances or stochastically characterized signals. There are two families of methods: those which act at the residual generation step (unknown input observers (Alcorta-Garcia, Frank, 1997), disturbance optimal decoupling (Chen *et al.*, 1993)...) and those which act at the residual interpretation step (statistical decision methods (Basseville, Nikiforov, 1993), fuzzy interpretation (Cassar, Staroswiecki, 1994), ...).

DX generally ignores the robustness problem, focusing on the use of high level of abstraction models, which are qualitative or symbolic. Also widely used in DX, interval models (also known as semi-qualitative models) are based on the assumption that uncertainties are bounded (Armengol et al., 2001; Loiez, 1997). These have been investigated for several years in the DX community as realizing a perfect compromise between precision and robustness; more recently, interval models have been considered in pure FDI approaches (Adrot *et al.*, 1999; Ploix *et al.*, 2000).

Moreover the decoupling methods (Chen *et al.*, 1993), (Staroswiecki *et al.*, 1993) proposed by FDI to make ARRs insensitive to unknown disturbances have no equivalent in DX.

[8] The Hamming distance represents the number of edges separating two vertices in the $\{0,1\}^n$ hypercube, where n is the number of ARRs.

VI.4 Logical soundness vs. structural robustness

As seen in section 4, the DX logical diagnosis theory does not make any kind of assumption about the faults *a priori*, which guarantees logically sound results. In the most general case, single as well as multiple faults are considered: a fault may be observable or not at the symptom level and multiple faults may as well compensate, i.e. being themselves not observable. When the application domain suggests specific assumptions, these are explicitly stated as additional axioms, for example the exoneration assumptions as defined in section IV.1

Conversely, the FDI approach is rather guided by structural robustness properties, i.e. the diagnosis results are valid in all the observation/fault space but a tiny region corresponding to fault cases discarded *a priori*, e.g. non detectable faults, which leave the SM satisfied. As pointed out in II.1.4, the single fault assumption is frequently adopted in many FDI applications, because it happens frequently that multiple faults have a very low probability, and not considering them drastically simplifies the computation. When this hypothesis is not realistic, one must anticipate how multiple faults combine their effects in the ARRs.

This is why comparing the DX and FDI approaches calls for the definition of all underlying assumptions since they are by default different and that they do not necessarily match in their formal definitions, e.g. ARR-exo and COMP-exo assumptions.

Also related is the fact that, since FDI has a pre-compilation approach, the faults and sensors are chosen so that the diagnosis results are not ambiguous. Residuals are then so designed that they are at least weakly isolating (Gertler, 1993), i.e. that every of the considered faults has a different signature. If this is not possible, the solution calls for the redesign of the instrumentation system (Carpentier, 1999). Conversely, in DX, the so-formulated theory allows one to generate all diagnosis candidates and proposes techniques to stamp them with probability or preference degrees. In practice, these degrees are used on line to generate only the most plausible candidates (highest probabilities, best explanation capacity, etc.).

VII Conclusion

The first goal of FDI was historically fault detection and associated decision procedures. Its main interest was to offer sophisticated techniques, such as observers and filters, so as to interpret observations to produce a set of symptoms (residuals). Nevertheless, the residuals can be designed in such a way that they are also informative from the fault localization point of view. DX approached the diagnosis problem the other way around, focusing on fault localization by pointing out the subsets of the system description that conflict with the observations. Our study proves that a significant part of the two theories fits into a common framework which allows a precise comparison. When they adopt the same hypotheses with respect to how faults manifest themselves and how many faults can occur simultaneously, FDI and DX views agree on diagnoses. This opens the possibility of a fruitful cooperation between these two diagnostic approaches.

Some points have been left out of this comparison. There is presently no equivalent in DX of the notion of unknown disturbance or noise. Conversely, DX makes a systematic use of fault models, whose counterpart in FDI can be found in assumptions about the additive or multiplicative disturbances that model the faults but always with respect to a correct behavior model. Fault models have been left out of the framework of the present paper. Temporal aspects of diagnosis, which are crucial in the state tracking problem, have not been approached neither. Further studies are needed to integrate these aspects, which would be beneficial to both communities.

VIII References

Adrot O., Maquin D., Ragot J. (1999), Fault detection with model parameter structured uncertainties, *European Control Conference ECC'99,* Karlsruhe, Germany, August 31-September 3, 1999.

Alcorta-Garcia E, Frank P.M. (1997), Deterministic non-linear observer-based approaches to fault diagnosis: a survey, *Control Engineering Practice* 5(5), p. 663-670.

Armengol J, Vehi J., Travé-Massuyès L., Sainz M.A. (2001), Application of multiple sliding time windows to fault detection based on interval models, *12th International Workshop on Principles of Diagnosis DX'01*, Via Lattea, Italy, March 7-9, 2001, p. .9-16.

Basseville M., Cordier M.-O. (1996), *Surveillance et diagnostic de systèmes dynamiques: approches complémentaires du traitement du signal et de l'intelligence artificielle*, Publication IRISA n° 1004.

Basseville M., Nikiforov I. (1993), *Detection of abrupt changes – Theory and Applications*, Information and System Sciences Serie, Prentice Hall, Englewood Cliffs, N. J.Berge C. (1975), *Théorie des graphes et ses applications*, Dunod.

Brusoni V., Console L., Terenziani P., Theseider Dupré D. (1998), A spectrum of definitions for temporal model-based diagnosis, *Artificial Intelligence* 102(1), p. 39-79.

Carpentier T. (1999), *Placement de capteurs pour la surveillance de processus complexes*, Doctorat de l'Université des Sciences et Technologies de Lille, May 11, 1999.

Cassar J.-Ph., Staroswiecki M. (1994), Advanced Design of the Decision Procedure in Failure Detection and Isolation Systems, *IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS'94*, Espoo, Finland.

Cauvin S., Cordier M.-O., Dousson C., Laborie P., Lévy F., Montmain J., Porcheron M., Servet I., Travé-Massuyès L. (1998), Monitoring and alarm interpretation in industrial environments, *AI Communications*.

Chen J., Patton R.J., Zhang H.Y. (1993), A multi-criteria optimization approach to the design of robust fault detection algorithms, *International Conference on Fault Diagnosis Tooldiag'93*, Toulouse, France.

CEP (1997), Control Engineering Practice, *Special volume on Supervision, fault detection, and diagnosis of technical systems*, Vol. 5(5).

Cordier M.-O., Dague P., Dumas M., Lévy F., Montmain J., Staroswiecki M., Travé-Massuyès L. (2000a), AI and Automatic control approaches of model-based diagnosis: links and underlying hypotheses, *4th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS 2000*, Budapest, Hungary, June, 14-16, 2000.

Cordier M.-O., Dague P., Dumas M., Lévy F., Montmain J., Staroswiecki M., Travé-Massuyès L. (2000b), A comparative analysis of AI and control theory approaches to model-based diagnosis, *14th European Conference on Artificial Intelligence ECAI-00*, Berlin, August 20-25, 2000, p. 136-140.

Dague P., Taillibert P. (1996), The monostable: a reference problem for analog diagnosis, *7th International Workshop on Principles of Diagnosis DX'96*, Val Morin, Québec, p. 79-87.

Davis R. (1984), Diagnostic Reasoning based on structure and behavior, *Artificial Intelligence* 24, p. 347-410.

De Kleer J. (1986)., An assumption-based TMS, *Artificial Intelligence*, 28(2), p. 127-162.

De Kleer J., Mackworth A., Reiter R. (1992), Characterizing diagnoses and systems, *Artificial Intelligence* 56(2-3), p. 197-222.

De Kleer J., Raiman O., Shirley M. (1991), One step lookahead is pretty good, *2nd International Workshop on Principles of Diagnosis DX'91*, Milan, p. 136-142. Also in *Readings in Model-Based Diagnosis*, Hamscher W., Console L., de Kleer J. (eds.), Morgan Kaufmann, 1992, p. 138-142.

De Kleer J., Williams B. C (1987), Diagnosing multiple faults, *Artificial Intelligence* 32(1), p. 97-130.

De Kleer J., Williams B. C. (1989), Diagnosis with behavioral modes, *11th International Joint Conference on Artificial Intelligence IJCAI-89*, Detroit, MI, p. 1324-1330.

Dubuisson B. (1990), *Diagnostic et reconnaissance des formes,* Traité des Nouvelles Technologies, Série Diagnostic et Maintenance, Hermès.

Frank P.M. (1996), Analytical and qualitative model-based fault diagnosis – A survey and some new results, *European Journal of Control,* Vol. 2, p. 6-28.

Gertler J.J. (1993), Analytical redundancy methods in fault detection and isolation, *International Conference on Fault Diagnosis Tooldiag'93,* Toulouse, France.

Gertler, J. (1998), *Fault detection and diagnosis in engineering systems*, Marcel Dekker Inc.

Greiner R., Smith B.A., Wilkerson R.W. (1989), A correction to the algorithm in Reiter's theory of diagnosis, *Artificial Intelligence* 47(1), p. 79-88.

Hamscher W., Console L., de Kleer J. (eds.) (1992), *Readings in Model-Based Diagnosis*, Morgan Kaufmann, San Mateo, CA.

Hodge W.V.D., Pedoe D. (1952), *Methods of Algebraic Geometry*, Cambridge University Press.

Iserman R. (1997), Supervision, fault detection and fault-diagnosis methods – An introduction, *Control Engineering Practice*, Vol. 5(5), p. 639-652.

Isermann, R. (1989). Process fault diagnosis based on process knowledge, *AIPAC'89, Advanced Information Processing in Automatic Control*, Nancy, France.

Krysander M., Nyberg M. (2002), Structural analysis utilizing MSS sets with application to a paper plant, *13th International Workshop on Principles of Diagnosis DX'02,* Semmering, Austria, May 2-4, 2002, p. 51-57.

Loiez E. (1997), *Contribution au diagnostic de systèmes analogiques, Modélisation par des bandes temporelles,* thèse de l'université des Sciences et Technologies de Lille, March 3rd, 1997.

Luong M., Maquin D., Huynh C.T., Ragot J. (1994), Observability, redundancy, reliability, and integrated design of measurement systems, *2nd IFAC Symposium on Intelligent Components for Control Applications SICICA'94*, Budapest, Hungary, June 8-10, 1994.

Patton R.J., Chen J. (1991), A review of parity space approaches to fault diagnosis, *IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS'91*, Baden-Baden.

Patton, R. J., Frank P., Clark R. (eds.) (2000), *Issues of Fault Diagnosis for Dynamic Systems*, Springer Verlag, 2000.

Peng Y., Reggia J. (1990), *Abductive inference models for diagnostic problem solving*, Springer-Verlag.

Ploix S., Adrot O., Ragot J. (2000), Bounding approach to the diagnosis of a class of uncertain static systems, *4th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS'2000,* Budapest, Hungary, June 14-16, 2000.

Pulido B., Alonso C. (2002), Possible conflicts, ARRs, and conflicts, *13th International Workshop on Principles of Diagnosis DX'02,* Semmering, Austria, May 2-4, 2002, p. 122-128.

Raiman O. (1992), The alibi principle, *Readings in Model-Based Diagnosis*, Hamscher W., Console L., de Kleer J. (eds.), Morgan Kaufmann, San Mateo, CA, p. 66-70.

Reiter R. (1987), A theory of diagnosis from first principles, *Artificial Intelligence* 32(1), p. 57-96.

Staroswiecki M., Declerck P. (1989), Analytical Redundancy in Non-linear Interconnected Systems by means of Structural Analysis, *IFAC / IMACS / IFORS Symposium on Advanced Information Processing in Automatic Control, AIPAC' 89*, Nancy, France.

Staroswiecki M., Cassar J.-P., Cocquempot V. (1993), Generation of optimal structured residuals in the parity space, *12th IFAC Word Congress*, Sydney, Australia, Vol. 5, p. 535-542.

Staroswiecki M. (1998), Fault Detection and Isolation Based on Numerical and Qualitative Models, Plenary lecture, *3rd IMEKO Int. Conf. on Acoustical and Vibratory Surveillance and Diagnostic Techniques*, Senlis, October 13-15, France.

Staroswiecki, M. (2001). Model based FDI: the control approach, *BRIDGE Workshop Notes*, Via Lattea, Italy.

Staroswiecki M., Comtet-Varga G. (2001), Analytical redundancy relations for fault detection and isolation in algebraic dynamic systems, *Automatic*a, 37(5), p. 687-699.

Travé-Massuyès L., Dague P., Guerrin F. (eds.) (1997), *Le raisonnement qualitatif pour les sciences de l'ingénieur*, Edition Hermès, Paris.

Travé-Massuyès L., Dague P. (1999), Etude dans le cadre du DE Autonomie, thème Diagnostic et Décision à bord. Rapport n°1 : *Approches pour le diagnostic*, Rapport CNES, Toulouse, France.