Artificial Intelligence 331 (2024) 104116

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint

Critical observations in model-based diagnosis

Cody James Christopher [a,*], Alban Grastien [b,c]

[a] Data61, CSIRO, Australia
[b] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[c] Humanising Machine Intelligence, The Australian National University, Australia

ARTICLE INFO

Keywords:
Explanation
Explainability
AI
Diagnosis
Model-based
Critical observation

ABSTRACT

In this paper, we address the problem of finding the part of the observations that is useful for the diagnosis. We define a *sub-observation* as an abstraction of the observations. We then argue that a sub-observation is *sufficient* if it allows a diagnoser to derive the same minimal diagnosis as the original observations; and we define *critical observations* as a maximally abstracted sufficient sub-observation. We show how to compute a critical observation, and discuss a number of algorithmic improvements that also shed light on the theory of critical observations. Finally, we illustrate this framework on both state-based and event-based observations.

1. Introduction

When a system that performs a useful task is subject to dysfunction, it likely needs some of its components repaired or replaced. The first step is then to perform a *diagnosis* based on *observations* made on the system such as sensor readings, event logs, etc. The result of this procedure, also called a *diagnosis*, is a set of possible *diagnosis candidates*, each of which being a description of a possible system status. While the diagnosis will inform the next steps (the decisions that will be made), we make no presumption here on how the diagnosis will be used.

We call the procedure that computes diagnoses a *diagnoser*. Most existing diagnosers only return a diagnosis and we argue as preliminary to this work that this is not satisfactory in a variety of situations. Indeed, when there is a cost associated with a diagnosis—for instance, because it indicates that an expensive component needs to be replaced—the *decision maker* in charge of addressing this diagnosis will often be reluctant to take action solely based on a black box recommendation. We would like to provide some form of *evidence* together with the diagnosis that could convince the decision maker of the validity of the diagnosis. This is even more salient when the diagnoser is unreliable, be it because the observability of the system is very limited or because the diagnoser has an imperfect description of the system (which can be the case for systems that are large, complex, and old such as power distribution networks). The decision maker can then decide whether the evidence is conclusive and prompts for action. Additionally, we posit that having some form of explanation of the diagnosis would help the decision maker understand the diagnosis itself, and also allow them to understand the limitations of the diagnoser.

This work is rooted in *consistency-based diagnosis* [1,2]. In this context, the diagnoser reasons on a *model*, i.e., a mathematical description of the system, to determine what could have happened in the system. If we assume that the model contains knowledge that is known to the decision maker (i.e., the decision maker knows how the system functions), then the observations can be seen as the evidence that supports the diagnosis: indeed, the model and the observations together entail the diagnosis. However, for

[*] Corresponding author.
E-mail addresses: cody.christopher@data61.csiro.au (C.J. Christopher), alban.grastien@cea.fr (A. Grastien).

https://doi.org/10.1016/j.artint.2024.104116
Received 1 February 2022; Received in revised form 7 February 2024; Accepted 15 March 2024

Available online 29 March 2024

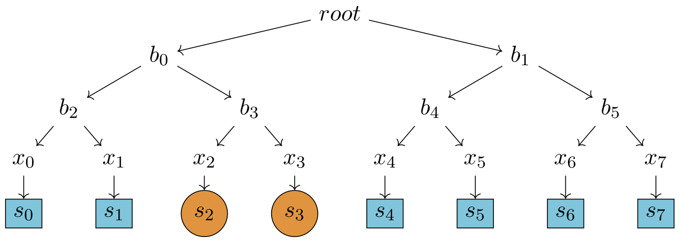

Fig. 1. A power network distributing power from *root* to consumers $x_0, \ldots, x_7$, each equipped with a sensor $s_0, \ldots, s_7$. Square nodes indicate power is detected; whereas circular nodes indicate it is not.

non-trivial systems such as large networks, the observations are generally overwhelmingly numerous for a decision maker to digest them in reasonable time, if at all. Instead, we would like to return only a small part of the observations: the *Critical Observations*.

We now illustrate the problem we are solving with a simplified example.

1.1. A motivating example

Consider a tree-shaped power network, such as given in Fig. 1,[1] where electricity flows from the root of the network through its buses b_j to its leaves x_i. When a component is faulty, the flow stops there and the sub-tree rooted at that location lose power. Each leaf x_i is equipped with a sensor s_i that reports whether the leaf is powered (for example, by measuring the voltage). We assume for simplicity that these sensors themselves cannot be faulty. We have here a *strong* model, i.e., a model that explains what happens when a fault occurs; however, the theory can accommodate a weak model instead. Together this means that a non-faulty component will, by definition, exonerate the entire path it lies on from itself back to the root.

The observation[2] in the specific instance of Fig. 1, with s_2 and s_3 observing no flow, yields five diagnosis candidates: $\{b_3\}$, $\{b_3, x_2\}$, $\{b_3, x_3\}$, $\{b_3, x_2, x_3\}$, $\{x_2, x_3\}$. Specifically, if all the components in a given candidate are faulty, it would explain the observation of the system. Due to the potentially large size of the diagnosis, diagnosers will usually return only the *minimal* candidates; those which are in some sense considered 'simpler' in the context. For this example, with respect to the subset relation, there are two such candidates: $\{b_3\}$ and $\{x_2, x_3\}$ – e.g. with $\{b_3\} \subset \{b_3, x_2\}$, we infer that b_3 failing alone is more likely than b_3 and x_2 both failing.

In this work we are not only interested in the diagnosis, but also in obtaining evidence for this diagnosis. For our running example, this can be reduced to a question about which observed values should be returned to the operator. Intuitively, it seems that at least both "abnormal" values (given by s_2 and s_3) should be kept, as both indicate that there is a faulty component. One might then infer that none of the other readings are useful by virtue of them being nominal. This is not the case: each powered sensor s_i shows, amongst other things, that the component it monitors, x_i, is non-faulty. This, in itself, is not useful information, given that we never presumed that these components were faulty to begin with. However, these sensors do provide useful information: s_7, for instance, indicates that at least the root, which was a suspect for the abnormal values of s_2 and s_3, must be working correctly. Sensors s_0 and s_1 are even more informative according to this criterion as they exonerate both the root and b_0. We thus conclude that $\{s_0, s_2, s_3\}$ and $\{s_1, s_2, s_3\}$ are both solutions to our problem.

1.2. Contribution and outline

Generalising from this example, we claim that the value in the observation is the fact that it allows one (the diagnoser or the decision maker) to derive the minimal diagnosis. All other information is either irrelevant (as when sensor s_7 shows that x_7 is not faulty, which we never suspected) or redundant (as when sensor s_1 shows that b_0 is not faulty when s_0 already proved it).

We define a *sub-observation* as an abstraction of the observation. We say that a sub-observation is *sufficient* (or, equivalently, refined enough) when it allows a diagnoser to derive the same minimal diagnosis as it would with the unabstracted observation. A sub-observation is then called *critical* if it is as abstract as possible while still remaining sufficient. We discuss sufficient conditions to guarantee the existence of a critical observation, how one can compute a critical observation, and how to define sub-observations in non-trivial settings, such as is present in the case of event-based observations. Our use of a model-based approach allows us to identify which part of the observation is useful in deriving the diagnosis, and which parts are either useless or redundant.

In Section 2, we recap a set of standard definitions for the problem of model-based diagnosis (or "diagnosis from first principles") that are at the core of this work. These definitions are general enough to capture circuits, continuous systems, discrete event systems, and hybrid systems. Section 3 presents our framework and, in particular, the definitions of sub-observations and critical observations. We discuss the computation of critical observations in Section 4. Section 5 demonstrates how the framework can be implemented

[1] Note that due to this example using state-based observations, it has some remarkable properties that do not hold in general; the framework eventually presented is more involved than this example may suggest is necessary, but this complexity allows us to address event-based observations in § 5.2.

[2] We use the singular form 'observation' to refer to the complete information returned by the sensors; once again, this is due to event-based observations in which the information is not just the list of observed events but may also include implicit information such as the fact that some events were not observed.

Table 1
Table of all symbols used in this paper.

Symbol	Definition
$\mathcal{M}$	model
$w \in \mathcal{M}$	run
$\mathbb{O}$	observation space
$\mathcal{O} \in \mathbb{O}$	observation
$\mathfrak{F}_{\mathbb{O}} : \mathcal{M} \longrightarrow \mathbb{O}$	observation function
$\mathbb{H}$	hypothesis space
$h \in \mathbb{H}$	hypothesis
$\mathfrak{F}_{\mathbb{H}} : \mathcal{M} \longrightarrow \mathbb{H}$	hypothesis function
$\preceq_{\mathbb{H}} \subseteq (\mathbb{H} \times \mathbb{H})$	preference relation over hypotheses
$\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$	diagnosis framework
$\mathcal{P} = \langle \mathcal{D}, \mathcal{O} \rangle$	diagnosis problem
Δ	diagnosis function and diagnosis
δ	diagnosis candidate
$\Delta_{\min}$	minimal diagnosis
Θ	sub-observation space
$\theta \in \Theta$	sub-observation
$\preceq \subseteq (\Theta \times \Theta)$	sub-observation abstraction relation
$sub : \mathbb{O} \longrightarrow \Theta$	sub-observation function
$\mathcal{S} = \langle \Theta, \preceq, sub \rangle$	sub-observation framework
$\psi : \Theta \longrightarrow 2^{\mathbb{O}}$	inverse of derivation function
$\mathcal{C}$	conflict
$\mathcal{X}$	observable conflict
∇	hypotheses not subsumed by candidates

for state-based and event-based observations. We discuss implementation issues in Section 6. Finally, we discuss related works in Section 7, and conclude in Section 8.

To clarify the paper, Table 1 lists most of the symbols used in this paper (ignoring those that are used in one specific place).

This is based on the work from the first author performed as part of a PhD program [3–7]. In this paper, we are able to give a more comprehensive justification for critical observations. All proofs of theorems, lemmas, and corollaries can be found in Appendix A. We also present a full framework, complete with a general definition of diagnosis (i.e., not limited to sets of faults as in [3,4] or single faults as in [5,6]) and the introduction of a sub-observation framework. We also provide additional improvements such as the utilisation of *observable conflicts* and a pruning technique in Section 6, along with associated insights to the theory of model-based diagnosis.

2. Diagnosis from first principles

Diagnosis is the problem of detecting and identifying faults in partially observable systems. In model-based diagnosis, this is done by comparing a given observation of the system with the observation predicted by a formal representation of said system, the eponymous “model”. In this section, we provide generic definitions inspired from the seminal papers in consistency-based diagnosis [1,2]. Our definition is meant to be generic enough to encompass a wide range of problems that includes both static and dynamic systems. The diagnosis definition requires essentially three components: the model, the observation, and the hypotheses.

The *system model*, $\mathcal{M}$, is a description of all behaviours of the target system. It is generally assumed that such models are complete (i.e., if the system exhibits a behaviour in reality, it is also contained in $\mathcal{M}$) unless stated otherwise. A *run*, w, of the system is a possible behaviour permitted by the model. With slight abuse of notation, we interpret the model as a set of possible behaviours and write $w \in \mathcal{M}$.

Example. In a static system, a model is generally comprised of i) a set of state variables that take values from a specified domain and ii) a set of constraints that indicate what assignments are possible. In our motivating example of Subsection 1.1, the model will include a Boolean variable for each component of the network that indicates whether this component is powered as well as a variable that indicates whether the component is faulty. The constraints explain how the electricity flows and that, *e.g.* if a component is faulty, then all the components in the sub-network rooted there are un-powered. A run is then a function that maps each variable to a value from its domain and that is consistent with the constraints.

In a dynamic system such as a discrete event system, the model is a state machine (automaton) and a run is then a trajectory on this automaton.

The actual behaviour of the system is unknown to the diagnoser. However, the diagnoser has access to some information about this behaviour, which can be provided by, *e.g.* sensor readings, message logging, etc. We call this information an *observation* $\mathcal{O}$ (often also called a *trace*). We write $\mathbb{O}$ for the set of possible observations, i.e., $\mathcal{O} \in \mathbb{O}$. The function $\mathfrak{F}_{\mathbb{O}} : \mathcal{M} \longrightarrow \mathbb{O}$ maps behaviours to what would be observed according to the model:

$$\mathfrak{F}_{\mathbb{O}}(w) = \mathcal{O}.$$

Example. In a static system modelled as a set of variables and a set of constraints, the observation is often the projection of the state onto a subset of 'observable variables'. In our motivating example, the observation is a mapping that indicates for each sensor whether power flows all the way to the sensor.

In a dynamic system modelled as a discrete event system, the observation is generally defined as the projection of the string of events in the run on the set of 'observable events'. If the model is timed, the observed events might be timed as well.

The stakeholder calling for a diagnosis is generally not interested in the specific system behaviour but in some more abstract information such as whether the system is faulty and, if so, in which part(s) of the system. A *diagnostic hypothesis*, h, is the description of a type of behaviour at a level that the stakeholder is interested in. Hypotheses are disjunctive, meaning that no two hypotheses can be true at any given time. The collection of all possible hypotheses is called the *hypothesis space*, $\mathbb{H}$. The function $\mathfrak{F}_{\mathbb{H}} : \mathcal{M} \longrightarrow \mathbb{H}$ maps behaviours to the hypothesis associated with that behaviour:

$$\mathfrak{F}_{\mathbb{H}}(w) = h.$$

Example. In a static system modelled as a set of variables and a set of constraints, the hypothesis is often the projection of the state onto a subset of 'fault variables', the subset of variables that describes the health of the system. In our motivating example, this is the set of variables that indicate whether a node of the network is faulty. A hypothesis can then be understood as a set of 'faults', where a fault indicates that a component is indeed faulty.

In a dynamic system such as a discrete event system, the model is often equipped with a subset of events called 'faulty events'; the hypothesis of a run is then generally the list of faulty events that occurred during the trajectory. There exist more complex definitions, *e.g.* where a hypothesis is the *sequence* of faulty events that occurred during the trajectory; where the order between the events is then considered relevant, and a single event could occur multiple times [8].

The hypothesis space is typically equipped with a partial order (having the properties of reflexivity, antisymmetry, and transitivity) relation denoted $\preceq_{\mathbb{H}}$, where $h \preceq_{\mathbb{H}} h'$ indicates that h is *preferred* to h'. The most preferred hypotheses are called *minimal*.

Example. When a hypothesis represents a set of faults, we often want to use the principle of parsimony and only consider hypotheses that include fewer faults. In this case, the preference relation $\preceq_{\mathbb{H}}$ is equivalent to the subset $\subseteq$ one, *i.e.* hypothesis h (understood as a set of faults) is preferred to h' if and only if h is a subset of h'. However, more complex definitions exist, *e.g.* where certain faults are assumed to be infinitely more unlikely than others: a hypothesis that does not include any such unlikely fault would then be preferred to one that does include some.

When a hypothesis $h \in \Sigma\star$ is a sequence of faults as can happen with discrete event systems, the preference relation is often the relation $\preceq_{\text{seq}}$ where $h \preceq_{\text{seq}} h'$ if h is a sub-sequence of h' (the faults in h appear in h' in the same order).

These elements together define a diagnosis framework:

Definition 1 *(Diagnosis framework).* A *diagnosis framework* is a tuple $\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$ where each element is as defined above.

A diagnosis problem is defined when a run w takes place in the system and the observation $\mathfrak{F}_{\mathbb{O}}(w)$ is available.

Definition 2 *(Diagnosis problem).* A *diagnosis problem*, $\mathcal{P}$, is a pair $\langle \mathcal{D}, \mathcal{O} \rangle$ where $\mathcal{D}$ is a diagnosis framework and $\mathcal{O}$ is an observation.

A problem $\mathcal{P}$ asks whether a given $\mathcal{O}$ is consistent with (or permitted by) $\mathcal{M}$ and under what conditions (i.e., whether faulty behaviour is necessary) $\mathcal{O}$ is exhibited. These conditions or behaviours are collected in those hypotheses drawn from $\mathbb{H}$.

Definition 3 *(Diagnosis).* Given a diagnosis framework $\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$ and a diagnosis problem $\mathcal{P} = \langle \mathcal{D}, \mathcal{O} \rangle$, the *diagnosis* $\Delta(\mathcal{P})$ is the collection of hypotheses that are consistent with the model and the system observation. These hypotheses are the *candidate hypotheses*, denoted δ:

$$\Delta(\mathcal{P}) = \left\{ \delta \in \mathbb{H} \;\middle|\; \exists w \in \mathcal{M}.\; \mathfrak{F}_{\mathbb{O}}(w) = \mathcal{O} \;\wedge\; \mathfrak{F}_{\mathbb{H}}(w) = \delta \right\}. \tag{1}$$

When clear from the context, the diagnosis framework or even the diagnosis problem will be omitted and the diagnosis will simply be denoted $\Delta(\mathcal{O})$ or Δ. We refer to the procedure that computes a diagnosis as a *diagnoser*.

Equation (1) computes exactly all the hypotheses that are consistent with the observation in a given problem and framework. This implies that i) if a candidate δ appears in Δ, then it is indeed possible, according to the model, that δ is the hypothesis of the system run; and ii) if a hypothesis δ' does not appear in Δ, then it is impossible according to the model that δ' is the hypothesis of the system run. We are interested in minimal candidates described below.

Definition 4 *(Minimal/preferred candidate hypothesis).* Given a diagnosis framework $\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$ and a diagnosis problem $\mathcal{P} = \langle \mathcal{D}, \mathcal{O} \rangle$, a candidate hypothesis, $\delta \in \Delta$, is *minimal* (or *most preferred*) if there is no other candidate $\delta' \in \Delta$ that pre-orders δ with respect to $\preceq_{\mathbb{H}}$:

$$\delta \text{ minimal } \iff \forall \delta' \in \Delta.\ \delta' \neq \delta\ \Rightarrow \delta' \npreceq_{\mathbb{H}} \delta.$$

This then establishes the minimal diagnosis of a problem:

Definition 5 *(Minimal diagnosis).* Given a diagnosis framework $\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$ and a diagnosis problem $\mathcal{P} = \langle \mathcal{D}, \mathcal{O} \rangle$, the *minimal diagnosis*, $\Delta_{\text{min}}(\mathcal{P})$, is the collection of minimal hypotheses in the diagnosis $\Delta(\mathcal{P})$:

$$\Delta_{\text{min}}(\mathcal{P}) = \{\delta \in \Delta(\mathcal{P})\ \mid \delta \text{ minimal in } \Delta(\mathcal{P})\,\}\,. \tag{2}$$

Again, when clear from the context, we replace $\Delta_{\text{min}}(\mathcal{P})$ with $\Delta_{\text{min}}(\mathcal{O})$ or Δ_{min}.

The minimal diagnosis can be viewed as the list of 'best' interpretations of the system observation in the sense that, according to the model, all diagnosis candidates are at least as bad as one of the minimal candidates. This is expressed in the lemma below, but first requires the following assumption.

We make the assumption that the hypothesis space $\langle \mathbb{H}, \preceq_{\mathbb{H}} \rangle$ is a *well-partial-order*, i.e., that any subset $H \subseteq \mathbb{H}$ of hypotheses has a finite, and non-empty, number of minimal elements:

$$\forall H \subseteq \mathbb{H}. \quad H \neq \emptyset \Rightarrow 0 < \left| \min_{\preceq_{\mathbb{H}}}(H) \right| < \infty$$

where $\min_{\preceq}(S) = \left\{ e \in S \mid \forall e' \in S.\ e' \preceq e \Rightarrow e = e' \right\}$ is the set of minimal elements of S according to $\preceq$. Most natural hypothesis spaces are well-partial-orders. This includes the powerset equipped with the subset relation: $\left\langle 2^F, \subseteq \right\rangle$, and the Kleene closure with the subsequence relation: $\left\langle \Sigma^\star, \preceq_{\text{seq}} \right\rangle$ [9].

However, some hypothesis spaces are not well-partial-orders. We take an example in which the system includes a single parameter, and the hypothesis represents the absolute discrepancy between the parameter's nominal value and its actual value. The hypothesis space is then $\mathbb{H} \overset{def}{=} \mathbf{Q}^+$, the set of positive rationals, and the preference relation is $\preceq_{\mathbb{H}} \overset{def}{=} \leq$ (less-or-equal-to) as we prefer hypotheses with a small discrepancy. In this case, the subset of hypotheses $H = (1, 2]$ between 1 (exclusive) and 2 (inclusive) has no minimal element.

Assumption 1. The hypothesis space $\langle \mathbb{H}, \preceq_{\mathbb{H}} \rangle$ is a well-partial-order.

Lemma 2.1. *Under Assumption 1, the minimal diagnosis* Δ_{min} *satisfies*

$$\forall \delta \in \Delta.\ \exists \delta' \in \Delta_{\text{min}}.\ \delta' \preceq \delta.$$

From this point on, we will no longer explicitly refer to Assumption 1. We interpret Lemma 2.1 as follows: given the minimal diagnosis Δ_{min}, we may be missing some diagnosis candidates $\delta \in \Delta \setminus \Delta_{\text{min}}$. However, these candidates are not minimal. Furthermore, Δ_{min} contains other candidates, such as δ', that are preferred over δ. In other words, it is acceptable to ignore candidates such as δ and to only return Δ_{min}. For these reasons, we will often content ourselves with the minimal diagnosis.

3. Critical observations

In this section we present the core contribution of this work, the theory of *Critical Observations* introduced by the authors in [4,6]. We present a generalised mathematical framework for the computation of critical observations, using the simple power network example of Fig. 1 to provide clarity to the concepts.

3.1. Sub-observations

A *sub-observation* θ is an object that represents some information about the system run, although the framework is general and does not specify what this information looks like. We write Θ for the set of possible sub-observations, i.e. $\theta \in \Theta$. In our example from Subsection 1.1, θ reports the readings from a subset of sensors.

A *sub-observation framework* is essentially a partially ordered set of sub-observations $\langle \Theta, \preceq \rangle$, which represent all the possible ways that any observation could be abstracted. The relation $\theta \preceq \theta'$ indicates that θ is more abstract (contains less information) than θ'. The sub-observation framework also specifies how these sub-observations relate to an observation: the function sub maps each observation $\mathcal{O}$ with the sub-observation $\theta = sub(\mathcal{O})$ that is equivalent to $\mathcal{O}$ in terms of the information it contains.

In our theory, the diagnoser will report a sub-observation θ with the meaning that the observation $\mathcal{O}$ (known by the diagnoser but assumed to be too information-rich to pass on to the user) satisfies $\theta \preceq sub(\mathcal{O})$. While the user will not know $\mathcal{O}$, θ will be chosen such that it contains enough information that the user will be able to derive a similar diagnosis as they would with $\mathcal{O}$; what 'similar' means in our context is explained in the rest of this section.

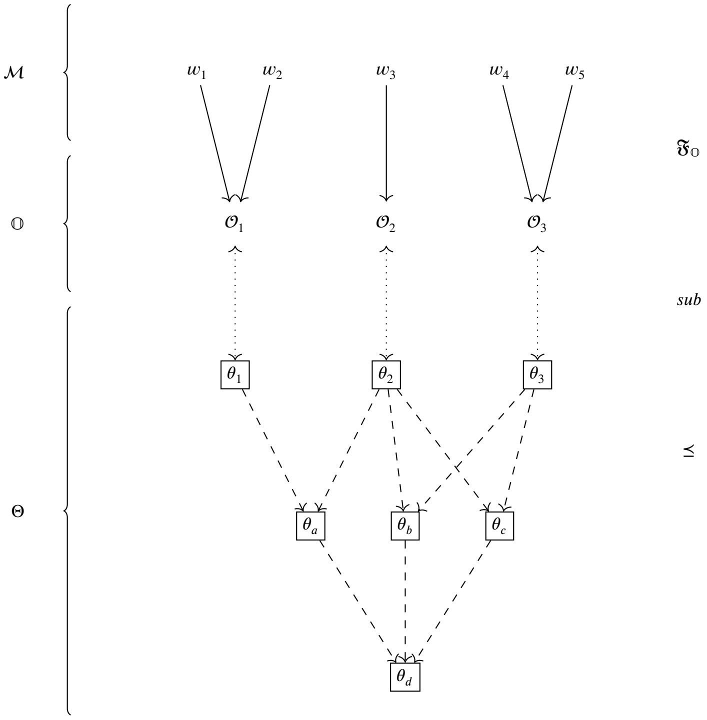

Fig. 2. Visualising the sub-observation space Θ. Plain arrows represent the observation function $\mathfrak{F}_{\mathbb{O}}$; dotted lines represent the *sub* function; dashed lines represent the abstraction relation $\preceq$.

An important requirement of our framework is that it should allow the diagnoser to report a sub-observation that contains precisely the same information as the actual observation. This is achieved through the following two assumptions. We first assume that *sub* maps every observation to a different sub-observation ($sub(\mathcal{O}_1) = sub(\mathcal{O}_2) \Rightarrow \mathcal{O}_1 = \mathcal{O}_2$). This is natural because we assume that *sub* is just a conversion from the observation space $\mathbb{O}$ to the sub-observation space Θ.

This needs to be pushed even further by way of the second assumption. We assume that the sub-observation $sub(\mathcal{O})$ is maximal (w.r.t. $\preceq$) amongst all sub-observations from $\mathbb{O}$; formally: $sub(\mathcal{O}) \preceq sub(\mathcal{O}') \Rightarrow \mathcal{O} = \mathcal{O}'$. The reason for this assumption is as follows. If there were two different observations $\mathcal{O}$ and $\mathcal{O}'$ with $sub(\mathcal{O}) \preceq sub(\mathcal{O}')$, then any sub-observation of $\mathcal{O}$ would also be a sub-observation of $\mathcal{O}'$. Hence, it would be impossible to report a sub-observation that only refers to $\mathcal{O}$.

Finally, we want to define the sub-observation framework separately from the diagnosis framework so that it can be studied independently. The framework is therefore defined over a set $\mathbb{O}'$ of observations that is a superset of $\mathbb{O}$. For a state-based framework for instance, $\mathbb{O}'$ is the set of all possible state-based observations for all systems and all sets of state variables.

Definition 6 *(Sub-observation framework).* A *sub-observation framework* is given as a tuple $\mathcal{S} = \langle \Theta, \preceq, sub \rangle$ such that

(a) Θ is a space of *sub-observations*,
(b) $\preceq$ is a partial order over Θ,
(c) $sub : \mathbb{O}' \longrightarrow \Theta$ is a function that maps observations $\mathcal{O} \in \mathbb{O}'$ with a sub-observation $sub(\mathcal{O})$.

Given a diagnosis framework $\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$ with $\mathbb{O} \subseteq \mathbb{O}'$, $\mathcal{S}$ is a *sub-observation framework of* $\mathcal{D}$ if for all observation $\mathcal{O}$, $sub(\mathcal{O})$ is maximal w.r.t. $\preceq$ amongst all sub-observations of $\mathbb{O}$, i.e.,

$$\forall \left\{\mathcal{O}, \mathcal{O}'\right\} \subseteq \mathbb{O}.\ \mathcal{O} \neq \mathcal{O}' \Rightarrow sub(\mathcal{O}) \npreceq sub(\mathcal{O}'). \tag{3}$$

When the relation $\theta \preceq \theta'$ holds, we say that θ is a *sub-observation of* θ'. If θ is different from θ', then we say that the relation is *strict* and write $\theta \prec \theta'$.

With these definitions in place, Fig. 2 serves to illustrate the shape of Θ, showing how sub-observations may relate. The possible behaviours of the model are represented at the top. These five behaviours can lead to three different observations $\mathcal{O}_1$ to $\mathcal{O}_3$ which can be computed through the $\mathfrak{F}_{\mathbb{O}}$ function. The *sub* function gives us three sub-observations for these observations, θ_1 to θ_3; the lines have arrows in both directions because it is possible to recover $\mathcal{O}$ from $sub(\mathcal{O})$. The abstraction relation is represented by the dashed lines. If we observe $\mathcal{O}_2$, we could then report the sub-observation θ_2, which is semantically equivalent to $\mathcal{O}_2$; we could instead report, for example, θ_a, which has had removed some of the information of the observation $\mathcal{O}_2$ in such a way that it becomes indistinguishable from the observation $\mathcal{O}_1$. Notice that we do not make any assumption on the abstraction relation beyond the fact that it is a partial order; in particular, this function is not necessarily a lattice, as illustrated by the fact that both θ_2 and θ_3 are supremum of θ_b and θ_c.

The diagnosis problem, however, is only defined for a specified observation and the associated diagnosis, and so we only ever care about the sub-observations that permit *at least* the observation in question. To make this explicit, we denote the sub-space of Θ rooted at $\theta = sub(\mathcal{O})$, the sub-observation that we start from, as $\Theta(\mathcal{O})$. Equivalently stated, this is the space of all sub-observations up to and including $sub(\mathcal{O})$ that permit the model to reproduce the behaviour that resulted in the observation $\mathcal{O}$:

$$\Theta(\mathcal{O}) \overset{def}{=} \left\{\theta' \in \Theta \;\middle|\; \theta' \preceq sub(\mathcal{O})\right\} \tag{4}$$

In the example of Fig. 2, the set $\Theta(\mathcal{O}_3)$ then equals $\left\{\theta_3, \theta_b, \theta_c, \theta_d\right\}$.

Stated in this way, we note that any sub-observation θ implicitly represents the set of all behaviours for which it is a more abstract form. We say formally that θ is *derivable* from some observation $\mathcal{O}$ if $\theta \preceq sub(\mathcal{O})$. To make this relationship mathematically explicit, consider that for any finite system run, eventually we will reach a sub-observation that could be derived from a different system run (in the extreme case, this is the empty sub-observation, which models all runs). We introduce $\psi(\cdot)$ as a convenient notation to express the set of all observations from which a given sub-observation may have derived:

$$\psi(\theta) = \left\{\mathcal{O}' \in \mathbb{O} \mid \theta \preceq sub(\mathcal{O}')\right\} \tag{5}$$

The construction of the precedence relation $\preceq$ as a partial order of the relative abstractedness of sub-observations has an important consequence. For any given sub-observation θ', the observations represented by this sub-observation are at least all those observations represented by all other sub-observations for which θ' is a more abstract form, according to the precedence relation. That is to say, explicitly:

$$\theta' \preceq \theta \;\Rightarrow\; \psi(\theta) \subseteq \psi(\theta') \tag{6}$$

Also, by equation (3) of Definition 6, we note the following property:

$$\psi(sub(\mathcal{O})) = \{\mathcal{O}\}\,.$$

Example. We illustrate the concept of sub-observations on the power network example of Fig. 1. In this example, an observation $\mathcal{O}$ is a function that indicates whether each sensor of the network is powered. The observation from the example is:

$$\mathcal{O} = \left\{s_0 \mapsto N, s_1 \mapsto N, s_2 \mapsto \boldsymbol{F}, s_3 \mapsto \boldsymbol{F}, s_4 \mapsto N, s_5 \mapsto N, s_6 \mapsto N, s_7 \mapsto N\right\}$$

where $\mathcal{O}(s_i) = N$ and $\mathcal{O}(s_i) = \boldsymbol{F}$ respectively indicate a normal observed value (the sensor s_i is powered) and a faulty one (unpowered). We define a sub-observation $\theta \preceq sub(\mathcal{O})$ as a set of pairs $\langle v, d\rangle$ where v is a sensor and $d = \mathcal{O}(v)$ is its observed value. Notice that we could define the sub-observation as a partial function similar to $\mathcal{O}$, but we chose this notation in order to make it easier to distinguish observations from sub-observations.

We introduce a specific notation to facilitate the discussion: for our specific observation, $\theta_{[S]}$ refers to the sub-observation in which the sensors mentioned in subscript are ignored. So the sub-observation $sub(\mathcal{O}) = \theta_{[]}$ is:

$$\left\{\langle s_0, N\rangle, \langle s_1, N\rangle, \langle s_2, \boldsymbol{F}\rangle, \langle s_3, \boldsymbol{F}\rangle, \langle s_4, N\rangle, \langle s_5, N\rangle, \langle s_6, N\rangle, \langle s_7, N\rangle\right\}$$

and the sub-observation $\theta_{[3,4]}$ is

$$\left\{\langle s_0, N\rangle, \langle s_1, N\rangle, \langle s_2, \boldsymbol{F}\rangle, \langle s_5, N\rangle, \langle s_6, N\rangle, \langle s_7, N\rangle\right\}.$$

Sub-observation $\theta_{[3,4]}$ represents all the observations that match the observations produced by the sensors s_0 to s_2 and s_5 to s_7; specifically the set $\psi\left(\theta_{[3,4]}\right)$ evaluates to:

$$\left\{\begin{array}{l}
\left\{s_0 \mapsto N, s_1 \mapsto N, s_2 \mapsto \boldsymbol{F}, \underline{s_3 \mapsto \boldsymbol{F}, s_4 \mapsto \boldsymbol{F}}, s_5 \mapsto N, s_6 \mapsto N, s_7 \mapsto N\right\} \\
\left\{s_0 \mapsto N, s_1 \mapsto N, s_2 \mapsto \boldsymbol{F}, \underline{s_3 \mapsto \boldsymbol{F}, s_4 \mapsto N}, s_5 \mapsto N, s_6 \mapsto N, s_7 \mapsto N\right\} \\
\left\{s_0 \mapsto N, s_1 \mapsto N, s_2 \mapsto \boldsymbol{F}, \underline{s_3 \mapsto N, s_4 \mapsto \boldsymbol{F}}, s_5 \mapsto N, s_6 \mapsto N, s_7 \mapsto N\right\} \\
\left\{s_0 \mapsto N, s_1 \mapsto N, s_2 \mapsto \boldsymbol{F}, \underline{s_3 \mapsto N, s_4 \mapsto N}, s_5 \mapsto N, s_6 \mapsto N, s_7 \mapsto N\right\}
\end{array}\right.$$

There are 2^8 sub-observations in $\Theta(\mathcal{O})$ since we can ignore any subset of the eight sensors.

3.2. Diagnosis sub-problems

Sub-observations provide us with a natural extension to diagnosis problems, which we call diagnosis *sub-problems*. A sub-problem takes the original model $\mathcal{M}$ that produced the observation $\mathcal{O}$, along with a sub-observation $\theta \in \Theta$ derived from $\mathcal{O}$. We treat sub-problems in much the same way as a proper diagnosis problem, in that we seek the minimal diagnosis that the sub-observation produces over the model. Formally:

Definition 7 *(Diagnosis sub-problem).* Given a diagnosis framework $\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$, a sub-observation framework of $\mathcal{D}$ – $\mathcal{S}$, and a sub-observation of $\mathcal{S}$ – θ, a *diagnosis sub-problem* of $\mathcal{P}$ is a tuple $\mathcal{P}' = \langle \mathcal{D}, \mathcal{S}, \theta \rangle$.

We note that the diagnosis sub-problem $\mathcal{P}'$ does not know the full observation $\mathcal{O}$.

We now extend the diagnosis symbols, Δ and $\Delta_{\min}$, to have semantics for sub-problems and sub-observations.

Definition 8 *(Sub-problem diagnosis).* Given a diagnosis framework $\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$ and a diagnosis sub-problem $\mathcal{P}' = \langle \mathcal{D}, \mathcal{S}, \theta \rangle$, the *diagnosis* of $\mathcal{P}'$ is:

$$\Delta(\mathcal{P}') \overset{def}{=} \left\{ \delta \in \mathbb{H} \;\middle|\; \exists w \in \mathcal{M}.\ \theta \preceq sub(\mathfrak{F}_{\mathbb{O}}(w)) \wedge \delta = \mathfrak{F}_{\mathbb{H}}(w) \right\}$$

and the *minimal diagnosis* is:

$$\Delta_{\min}(\mathcal{P}') \overset{def}{=} \left\{ \delta \in \Delta(\theta) \;\middle|\; \nexists \delta' \in \Delta(\theta)\ .\ \delta' \prec_{\mathbb{H}} \delta \right\}.$$

We use both flexibly, and typically leave the model, hypothesis space, observation function, and hypothesis function as implicit (due to their fixed nature), for example:

$$\Delta(\theta) \overset{def}{=} \Delta(\langle \mathcal{D}, \mathcal{S}, \theta \rangle),$$

$$\Delta_{\min}(\theta) \overset{def}{=} \Delta_{\min}(\langle \mathcal{D}, \mathcal{S}, \theta \rangle).$$

Similar to the definition of diagnosis of an observation, the diagnosis of a sub-observation comprises exactly all hypotheses that are consistent with the input knowledge, here the model and θ. In particular, if δ appears in the diagnosis $\Delta(\theta)$, then there exists a system run w such that:

(a) the observation of w is compatible with θ (formally, $\theta \preceq \mathfrak{F}_{\mathbb{O}}(w)$) and
(b) the hypothesis associated with w is δ ($\mathfrak{F}_{\mathbb{H}}(w) = \delta$).

Conversely, if δ does not appear in $\Delta(\theta)$, then there is no such run.

It is possible to reformulate the diagnosis of a sub-observation as the union of the diagnoses $\Delta(\mathcal{O}')$ of all observations $\mathcal{O}'$ that θ is derivable from.

Lemma 3.1. *Let θ be a sub-observation. The following property holds:*

$$\Delta(\theta) = \bigcup_{\mathcal{O}' \in \psi(\theta)} \Delta(\mathcal{O}').$$

This has several consequences. First, we know from Definition 6 that $\psi(sub(\mathcal{O}))$ equals precisely $\{\mathcal{O}\}$, which proves the following corollary:

Corollary 3.2. *The following property holds:*

$$\Delta(\mathcal{O}) = \Delta(sub(\mathcal{O})).$$

A second consequence is that considering increasingly abstract sub-observations leads to larger and larger diagnoses. This is due to the fact that $\theta' \preceq \theta$ implies $\psi(\theta) \subseteq \psi(\theta')$.

Corollary 3.3. *Let θ and θ' be two sub-observations. The following property holds:*

$$\theta' \preceq \theta \Rightarrow \Delta(\theta) \subseteq \Delta(\theta').$$

Therefore, given a problem defined by an observation $\mathcal{O}$ and a sub-problem given by sub-observation θ such that $\theta \preceq sub(\mathcal{O})$ the diagnosis of the sub-problem contains at least those candidates of the original problem, $\Delta(\mathcal{O}) \subseteq \Delta(\theta)$.

Example. Returning to Fig. 1, let us consider the original state of the problem and the initial sub-observation:

$$sub(\mathcal{O}) = \theta_{[]} = \left\{ s_0 \mapsto N, s_1 \mapsto N, s_2 \mapsto \boldsymbol{F}, s_3 \mapsto \boldsymbol{F}, s_4 \mapsto N, s_5 \mapsto N, s_6 \mapsto N, s_7 \mapsto N \right\}$$

Table 2
Minimal diagnosis for some sub-observations of Fig. 1, those matching the original diagnoses in bold.

Sub-observation θ	$\Delta_{\min}(\theta)$
$\theta_{[]}$	$\{b_3\}, \{x_2, x_3\}$
$\theta_{[2]}$	$\{b_3\}, \{x_3\}$
$\theta_{[3]}$	$\{b_3\}, \{x_2\}$
$\theta_{[3,4]}$	$\{b_3\}, \{x_2\}$
$\theta_{[0]}, \theta_{[1]}, \theta_{[4]}, \dots, \theta_{[7]}$	$\boldsymbol{\{b_3\}, \{x_2, x_3\}}$
$\theta_{[0,1]}$	$\{b_0\}, \{b_3\}, \{x_2, x_3\}$
$\theta_{[0,4-7]}$	$\boldsymbol{\{b_3\}, \{x_2, x_3\}}$
$\theta_{[1,4-7]}$	$\boldsymbol{\{b_3\}, \{x_2, x_3\}}$
$\theta_{[0,1,4-7]}$	$\{root\}, \{b_0\}, \{b_3\}, \{x_2, x_3\}$
$\theta_{[0,1,2]}$	$\{b_0\}, \{b_3\}, \{x_3\}$
$\theta_{[0,1,3]}$	$\{b_0\}, \{b_3\}, \{x_2\}$
$\theta_{[2,3]}$	$\emptyset$
$\theta_{[0,1,2,3]}$	$\emptyset$

Consider $\theta_{[7]}$ – the diagnosis of this sub-observation is:

$$\begin{aligned}
\Delta(\theta_{[7]}) &= \Delta(\mathcal{O}) \cup \Delta(\mathcal{O}') \\
&= \{\{b_3\}, \{b_3, x_2\}, \{b_3, x_3\}, \{b_3, x_2, x_3\}, \{x_2, x_3\}\} \\
&\quad \cup \{\{b_3, x_7\}, \{b_3, x_2, x_7\}, \{b_3, x_3, x_7\}, \{b_3, x_2, x_3, x_7\}, \{x_2, x_3, x_7\}\} \\
\text{s.t. } \mathcal{O}' &= \{s_0 \mapsto N, s_1 \mapsto N, s_2 \mapsto \boldsymbol{F}, s_3 \mapsto \boldsymbol{F}, s_4 \mapsto N, \dots, s_6 \mapsto N, s_7 \mapsto \boldsymbol{F}\}
\end{aligned}$$

The diagnosis of the sub-observation $\theta_{[7]}$ has a very natural interpretation: it is the same as the diagnosis for the complete observation with the difference that we no longer know that x_7 is non-faulty. Notice that $\Delta(\theta_{[7]})$ is indeed a superset of $\Delta(\theta_{[]})$. The minimal diagnosis of $\theta_{[7]}$ is then $\{\{b_3\}, \{x_2, x_3\}\}$, i.e., precisely the same as $\theta_{[]}$. This is essential to our theory, as we describe formally in the next subsection: while $\theta_{[7]}$ is strictly less informative than the original observation, since it loses the information that s_7 is not faulty, we consider that $\theta_{[7]}$ has the same value from a diagnostic standpoint as it allows the observer to infer the same minimal diagnosis. From an explanatory standpoint, it is even better since it was stripped of some irrelevant information.

We provide the minimal diagnosis on this example for a range of sub-observations in Table 2. These minimal diagnoses can also be retrieved using the following rule:

For sub-observation θ_S,

(a) if S includes both s_2 and s_3, then $\Delta_{\min}(\theta_S) = \{\emptyset\}$;
(b) otherwise, if S includes s_0, s_1, and s_3 to s_7,
  - if S includes s_2, then $\Delta_{\min}(\theta_S) = \{\{root\}, \{b_0\}, \{b_3\}, \{x_3\}\}$;
  - otherwise, if S includes s_3, then $\Delta_{\min}(\theta_S) = \{\{root\}, \{b_0\}, \{b_3\}, \{x_2\}\}$;
  - otherwise, $\Delta_{\min}(\theta_S) = \{\{root\}, \{b_0\}, \{b_3\}, \{x_2, x_3\}\}$;

(c) otherwise, if S includes both s_0 and s_1,
  - if S includes s_2, then $\Delta_{\min}(\theta_S) = \{\{b_0\}, \{b_3\}, \{x_3\}\}$;
  - otherwise, if S includes s_3, then $\Delta_{\min}(\theta_S) = \{\{b_0\}, \{b_3\}, \{x_2\}\}$;
  - otherwise, $\Delta_{\min}(\theta_S) = \{\{b_0\}, \{b_3\}, \{x_2, x_3\}\}$;

(d) otherwise,
  - if S includes s_2, then $\Delta_{\min}(\theta_S) = \{\{b_3\}, \{x_3\}\}$;
  - otherwise, if S includes s_3, then $\Delta_{\min}(\theta_S) = \{\{b_3\}, \{x_2\}\}$;
  - otherwise, $\Delta_{\min}(\theta_S) = \{\{b_3\}, \{x_2, x_3\}\}$.

3.3. Sub-observations and minimal diagnoses

Thus far, we have only explored how sub-observations relate to the diagnosis of the problem $\mathcal{P}$ that they abstract from. However, as the last examination of the running example suggests, we are more concerned with how sub-observations relate specifically to the *minimal* diagnosis. Most diagnostic settings appeal to the Principle of Parsimony, and consequently consider the simplest hypotheses as the most likely, where simplicity is given by the chosen definition of minimality.

As given by Corollary 3.3, sub-observations can only ever produce more hypotheses when compared to those produced by the original observation, but this says little regarding how the minimality of that set of candidates might be affected. Should the minimal diagnosis change, we would concede that the explanatory power of that sub-observation has diminished, as it disagrees with the original diagnosis. In this subsection we explicitly characterise the relationship between a given observation and the sub-observations of that observation with respect to the property we wish to keep static: the minimal diagnoses.

In the following lemma, we show that any candidate δ for the original observation $\mathcal{O}$ that is minimal for a sub-observation $\theta \in \Theta(\mathcal{O})$ must then also be minimal for this observation $\mathcal{O}$. Keeping in mind that the minimality of a diagnosis provides a way of distinguishing whether one candidate is better than another, this lemma has an important consequence: if a sub-observation allows for the claim that no strictly better candidate than δ exists (w.r.t. the preference relation), then that same claim is also allowed for the original observation. In other words, we show that using a sub-observation is a legitimate way of proving the minimality of a diagnosis. Thus, such a sub-observation is at least as useful to an operator as the original observation, even though some extraneous information has been removed.

Lemma 3.4. *Given observation $\mathcal{O}$ and sub-observation $\theta \in \Theta(\mathcal{O})$, where $\theta \preceq sub(\mathcal{O})$, the following property holds:*

$$\left(\Delta_{\min}(\theta) \cap \Delta(\mathcal{O})\right) \subseteq \Delta_{\min}(\mathcal{O}).$$

The next lemma improves this result further by showing that there can be no minimal candidates other than those implied by the sub-problem, *if* all those minimal candidates are candidates for the original problem.

Lemma 3.5. *Given observation $\mathcal{O}$ and sub-observation θ such that $\theta \in \Theta(\mathcal{O})$, where $\theta \preceq sub(\mathcal{O})$, then the following property holds:*

$$\left(\Delta_{\min}(\theta) \subseteq \Delta(\mathcal{O})\right) \Rightarrow \left(\Delta_{\min}(\mathcal{O}) \subseteq \Delta_{\min}(\theta)\right).$$

Having now established that if all minimal candidates of θ are valid candidates of $\mathcal{O}$, then all minimal candidates of $\mathcal{O}$ are also minimal candidates of θ, we can combine Lemma 3.4 with Lemma 3.5 to prove that if all minimal candidates of θ are candidates of $\mathcal{O}$, then these minimal candidates are identical:

Corollary 3.6. *Given observation $\mathcal{O}$ and a sub-observation $\theta \in \Theta(\mathcal{O})$, with $\theta \preceq sub(\mathcal{O})$, then the following property holds:*

$$\left(\Delta_{\min}(\theta) \subseteq \Delta(\mathcal{O})\right) \Rightarrow \left(\Delta_{\min}(\mathcal{O}) = \Delta_{\min}(\theta)\right).$$

This allows us to conclude that the sub-observation being considered, should it satisfy these relationships, has the same explanatory power as the original observation. Consequently, if a sub-observation determines that a system is experiencing all the faults from at least one member of the set of hypotheses $\left\{\delta_1, \ldots, \delta_k\right\}$, then this claim must still be true with the original observation. In other words, we have shown that there are no alternative explanations for the observations outside those of $\Delta_{\min}(\theta)$.

There are some caveats, however. Whilst a sub-observation is sufficient to disprove the validity of a diagnosis (Lemma 3.3), it is not sufficient in general to, conversely, *prove* validity. Indeed, notice that both Lemma 3.4 and Lemma 3.5 apply only to the minimal candidates of $\mathcal{P}'$ that are also candidates of $\mathcal{P}$. This is relevant in the case parsimony does not hold, and the actual hypothesis could indeed actually be worse (w.r.t. the preference relation) than one of the minimal diagnosis.

Furthermore, we have only shown these results to hold in a consistency-based diagnostic framework, but not in a probabilistic one. Logical consistency-based diagnosis enjoys the monotonicity of entailment — whereby adding logical statements (state-based observations in this context) cannot make an already invalid explanation valid ($\Gamma \vDash \bot \Rightarrow \Gamma \wedge \alpha \vDash \bot$) — that probabilistic frameworks unfortunately do not, as an explanation which was unlikely (compared to other explanations) can suddenly become highly probable if added observations support this explanation and contradict others. The validity of a diagnosis in such a framework is established by a probability threshold, as opposed to satisfiability in the context of consistency-based diagnosis.

Our last corollary relates the minimal diagnoses of increasingly abstract sub-observations.

Corollary 3.7. *Let θ_1, θ_2, and θ_3 be sub-observations such that $\theta_3 \preceq \theta_2 \preceq \theta_1$ holds. The following property holds:*

$$\left(\Delta_{\min}(\theta_1) = \Delta_{\min}(\theta_3)\right) \;\Rightarrow\; \left(\Delta_{\min}(\theta_1) = \Delta_{\min}(\theta_2) = \Delta_{\min}(\theta_3)\right).$$

3.4. Sufficiency & criticality

Given our framework of sub-observations and given the theoretical results that this framework enjoys, we can now concentrate on finding "good" sub-observations. The quality of a sub-observation is defined over two dimensions: i) it should allow a user to draw conclusions about the behaviour of the system, and ii) it should be as abstract as possible, i.e., be devoid of unnecessary information.

As we have shown earlier, a sub-observation cannot improve the diagnosis in the sense that it would allow one to remove incorrect candidates. We also stressed that the precise diagnosis is not an interesting object in itself, and that a decision maker in charge of monitoring the system is generally only interested in the minimal diagnosis.

As such, given a diagnosis problem $\mathcal{P}$, we want to return a sub-observation θ such that $\Delta_{\min}(\theta)$ equals $\Delta_{\min}(\mathcal{O})$.

Formally, we say that a sub-observation θ is *sufficient* if the sub-problem it defines produces the same minimal diagnosis as the original problem from which it derives:

Definition 9 *(Sufficiency).* Given a problem $\mathcal{P}$ and a set of hypotheses for $\mathcal{P}$, $D \subseteq \mathbb{H}$, a sub-observation $\theta \preceq sub(\mathcal{O})$ is *sufficient* to explain D if $\Delta_{\min}(\theta) \subseteq D$.

Given an observation $\mathcal{O}$, a sub-observation $\theta \preceq sub(\mathcal{O})$ is *sufficient* to explain $\mathcal{O}$ if $\Delta_{\min}(\theta) = \Delta_{\min}(\mathcal{O})$.

We now define the notion of a most sufficient sub-observation, what we call *critical*:

Definition 10 *(Criticality).* Given a problem $\mathcal{P}$ and a set of hypotheses for $\mathcal{P}$, $D \subseteq \mathbb{H}$, a sub-observation $\theta \preceq sub(\mathcal{O})$ is *critical* to explain D if it is sufficient to explain D and no sub-observation is sufficient to explain D.

Given an observation $\mathcal{O}$, a sub-observation $\theta \preceq sub(\mathcal{O})$ is *critical* to explain $\mathcal{O}$ if it is sufficient to explain $\mathcal{O}$ and no sub-observation of θ is sufficient to explain $\mathcal{O}$.

A critical sub-observation (more simply called a critical observation) is therefore a sufficient sub-observation that cannot be abstracted further without compromising explanatory power with respect to the precision of the original diagnosis. It is important to note that since the ordering relation only needs to be partial, there may be several possible critical observations for any given observation. The quality of any specific result as compared to any other is dependent on implementation decisions, such as the method chosen to explore Θ.

While we defined sufficiency for both a set of hypotheses and a specific observation, we concentrate on the second one in this work.

Lemma 3.8 *(Monotonicity of sufficiency). Given an observation $\mathcal{O}$ and two sub-observations θ_1, θ_2 such that $\theta_1 \preceq \theta_2 \preceq sub(\mathcal{O})$, if θ_1 is sufficient for $\mathcal{O}$, then so is θ_2.*

Example. Returning to the example in Fig. 1 with the sub-observations we previously considered (see Table 2), we are now equipped to determine whether these sub-observations are sufficient, critical, or neither.

Consider again $\theta_{[7]}$ — we find that diagnosing the sub-problem induced by this sub-observation produces the same minimal diagnosis (noting the original diagnosis with the addition of x_7, *e.g.* $\{x_2, x_3, x_7\}$, are also now considered valid hypotheses, but are not minimal). Therefore, according to Definition 9, we would state that θ is *sufficient*. We could not however demonstrate criticality, as we know that other sub-observations that are strict subsets of this one are also sufficient.

To contrast, we consider the case of $\theta_{[2]}$. Here, notice that we now introduce a new minimal hypothesis, $\delta = \{x_3\}$. Since $\Delta_{\min}(\theta_{[2]}) \neq \Delta_{\min}(\mathcal{O})$, this is an example of an insufficient sub-observation.

Lastly, consider $\theta_{[1,4,5,6,7]} = \{\langle s_0, N\rangle, \langle s_2, \boldsymbol{F}\rangle, \langle s_3, \boldsymbol{F}\rangle\}$, and the similar one given by $\theta_{[0,4,5,6,7]}$. According to Definition 10, these both satisfy the definition of *sufficiency*. We verify this by checking the minimal diagnosis, which we find as $\Delta_{\min}(\theta) = \Delta_{\min}(\mathcal{O}) = \{\{x_2, x_3\}, \{b_3\}\}$. Notice that many other non-minimal hypotheses may also be consistent with these sub-observations — specifically those found by combining every other suspect component (aside from those exonerated by the one remaining nominal reading) with the existing candidates.

Both of these are also *critical*. Each only has three possible removals, however removing any of the remaining observations alters the minimal diagnosis, as indicated in Table 2. This is consistent with the intuitions we developed for this model, as removing either of the remaining faulty observations ($\{s_2, s_3\}$) results in the loss of too much information. Similarly, removing the remaining nominal observation also alters the minimal diagnosis to include b_0, for example. As such, both $\theta_{[1,4,5,6,7]}$ and $\theta_{[0,4,5,6,7]}$ are critical for this example, and indeed, are the only critical sub-observations.

4. Computing critical sub-observations

With the notion of criticality defined for sub-observations, we now provide an algorithm to compute a critical sub-observation for a given diagnosis problem. Throughout this section we discuss other conditions necessary to develop a complete, correct, and terminating procedure prior to proving that the algorithm eventually provided is indeed all of these. To achieve this, a stable method to traverse the search space ($\Theta(\mathcal{O})$) is required, which we first provide before introducing the aforementioned algorithm.

4.1. Children of sub-observations

To define a search algorithm for critical observations, we need to formally provide a familial relation of Θ such that any search can be appropriately directed. For a given sub-observation θ we define the *children* of θ, such that there are no intervening levels of abstractions between those children and θ — that is, θ' and θ have no sub-observation *between* them:

Definition 11 *(Children).* A *child* of sub-observation θ is a strict sub-observation of θ such that it is the sub-observation of no other sub-observation of θ.

$$\theta' \in children(\theta) \Longleftrightarrow \left(\theta' \prec \theta\right) \wedge \left(\nexists \theta'' \in \Theta \,.\, \theta' \prec \theta'' \prec \theta\right). \tag{7}$$

We extend this naturally to children of sets of sub-observations, denoted $\mathfrak{S}$:

$$children(\mathfrak{S}) \overset{def}{=} \bigcup_{\theta \in \mathfrak{S}} (children(\theta)) \tag{8}$$

This definition of the children of a sub-observation provides a natural way to navigate the sub-space structure, with which we use to locate a critical observation. The definition of $children(\cdot)$ is also useful to characterise critical sub-observations as we show later in this section.

For the next set of results, we require an additional assumption on the sub-observation framework. This assumption specifies that the set $\Theta(\mathcal{O})$ of sub-observations that are derivable from any observation $\mathcal{O}$ should be finite.

The first reason why this assumption is required is that we will need to explore $\Theta(\mathcal{O})$ in order to find a critical observation. It should be clear that if $\Theta(\mathcal{O})$ is infinitely large, then a generic algorithm may never terminate.

The second issue arises if $\Theta(\mathcal{O})$ contains infinite descending sequences or, stated differently, if $\Theta(\mathcal{O})$ is not *well ordered*. That is, there could be an infinite sequence of sub-observations $\theta_1, \theta_2, \dots$ such that $\theta_{i+1} \prec \theta_i$ holds for all indices i. For sequences of this nature, there is no definite minimal element. Recall that a critical observation is a minimal element (w.r.t. $\preceq$) from the set of sufficient sub-observations. Therefore, if $\Theta(\mathcal{O})$ is infinitely large, there might be no critical observation.

Finally, a last issue arises if $\Theta(\mathcal{O})$ contains infinite *ascending* sequences, i.e., there exists an infinite sequence of sub-observations $\theta_1, \theta_2, \dots$ such that $\theta_i \prec \theta_{i+1}$ holds for all index i (noting that the existence of an infinite descending sequence does not imply the existence of an infinite ascending one, and vice versa). Remember that $\theta = sub(\mathcal{O})$ is the unique maximal sub-observation of $\Theta(\mathcal{O})$, i.e., $\theta_i \prec \theta$ for all index i. In the set $\left\{\theta_1, \theta_2, \dots\right\} \cup \{\theta\}$, θ has no child. This means that we cannot use the $children(\cdot)$ relation to traverse during search from θ to the critical observation(s).

We give a simple example of an infinite sub-observation space. Consider a system with a single sensor that reports a temperature t. Furthermore, consider that a sub-observation is an interval $[l, u]$ over the rational numbers such that: $sub(t) = [t, t]$ and $[l, u] \preceq [l', u']$ if and only if $l \leq l'$ and $u \geq u'$. Sub-observation $[l, u]$ indicates that the temperature is between l and u, inclusive. One first issue here could be that u is unbounded (l is bounded by 0 kelvin). We will assume that u is bounded, i.e., the sub-observation space only allows intervals with $u \leq U$ for some constant U. Even with this assumption, the sub-observation space is not well-founded as it contains infinite descending sequences such as this one:

$$[0, 9] \succeq [0, 9.9] \succeq [0, 9.99] \succeq [0, 9.999] \succeq \dots$$

With this type of sub-observation framework, there may be no critical observation for a given problem. In this situation, one should probably discretise the sub-observation space as going through an infinite sequence of infinitely small abstractions is not useful given our goal.

Assumption 2. The set $\Theta(\mathcal{O})$ of sub-observations derivable from any observation $\mathcal{O}$ has finite cardinality.

Assumption 2 gives us two main benefits. First, that any sub-observation derivable from $sub(\mathcal{O})$ can be reached from $sub(\mathcal{O})$ by taking *atomic steps* in the sub-observation space, where an atomic step consists in moving from a sub-observation to one of its children. This will allow us to design an algorithm for finding a minimal sub-observation. Second, that a sufficient sub-observation is critical if and only if its children are not sufficient. This is summarised in the following two lemmas:

Lemma 4.1. *Let θ and θ' be two sub-observations such that $\theta \prec \theta'$ holds. Under Assumption 2, there exists a sequence of sub-observations $\theta_0, \dots, \theta_k$ where $\theta_0 = \theta$, $\theta_k = \theta'$, and for all index $i \in \{0, \dots, k-1\}$, θ_i is a child of θ_{i+1}.*

Lemma 4.2. *Under Assumption 2, given a sufficient $\theta \preceq sub(\mathcal{O})$ such that for all $\theta' \in children(\theta)$, θ' is insufficient, then θ is critical for $\mathcal{O}$.*

We can now entirely visualise the subspace of the running example. In Fig. 3 we illustrate the entirety of $\Theta(\mathcal{O})$, taking some representational liberties as the full space with all precedence relationships is too large to explicitly show. Each level of this graph is a different stage of abstraction, with all nodes at the same level considered to be equally abstracted with respect to $sub(\mathcal{O})$. The depth from the root of graph indicates increasing abstraction as we descend until we reach the empty observation at the bottom. The notation used at each level is akin to a regular expression, indicating the patterns common to groups of sub-observations at that level. For example, $[(0|1), (4-7)\{2\}]$ indicates that either of the observations produced by s_0 or s_1 have been removed (but not both) as well as two of those out of $\left\{s_4, \dots, s_7\right\}$. A $*$ is wild, so in $[* \{6\}]$, for example, this means any six of the observations has been removed.

At each level, solid lines indicate which nodes are the direct children of nodes above, whilst dotted lines indicate potential children (necessary as a result of the representation). Formally, a solid line from the list of sub-observations $\mathfrak{S}$ to $\mathfrak{S}'$ indicates:

$$\forall \theta' \in \mathfrak{S}'.\ \exists \theta \in \mathfrak{S}.\ \theta' \in children(\theta)$$

and a dotted line indicates

$$\exists \theta' \in \mathfrak{S}'.\ \exists \theta \in \mathfrak{S}.\ \theta' \in children(\theta).$$

Note that for any two connected nodes, the more abstracted node will by definition be a sub-observation of the less abstracted node. A thick and solid line bisects the graph, showing the boundary of sufficiency. Importantly, note that there is no edge that crosses this boundary from insufficient to sufficient. To reiterate, once sufficiency has been lost (i.e., necessary data has been abstracted), it cannot be regained by further abstraction (Lemma 4.2).

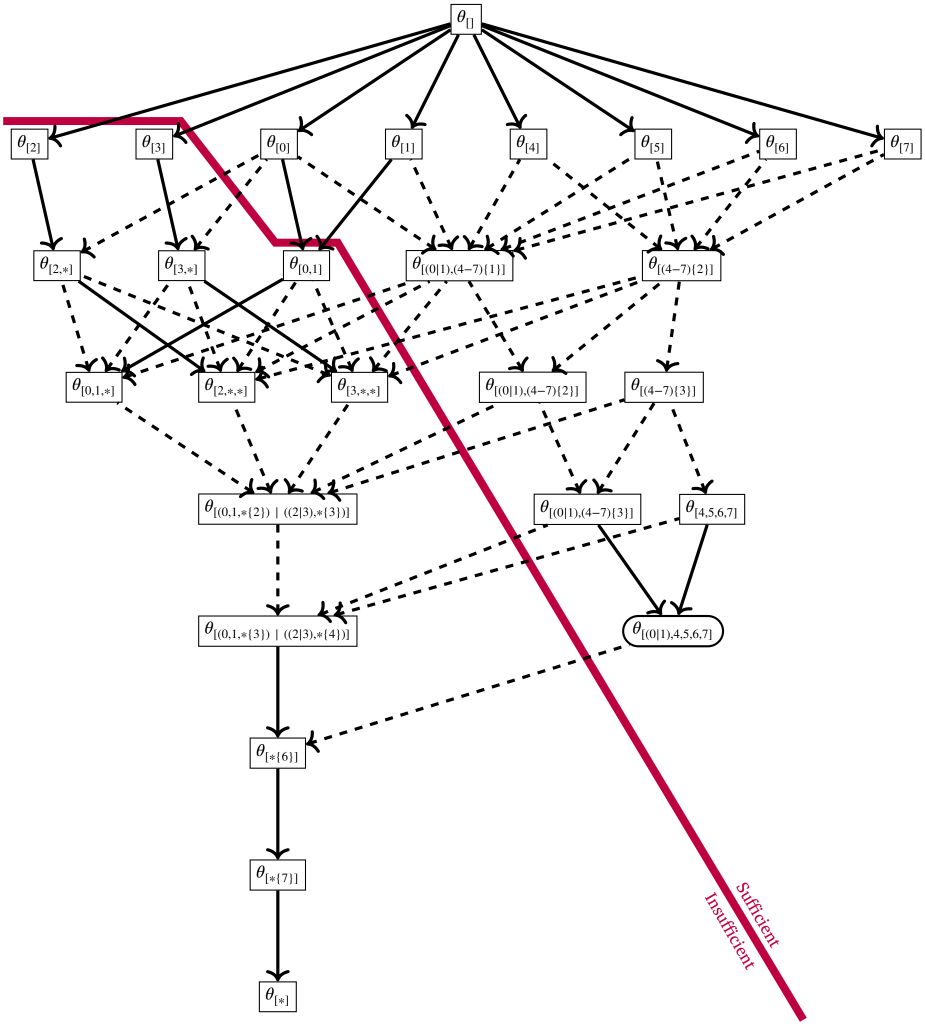

Fig. 3. Illustration of $\Theta(\mathcal{O})$ with partitioning line indicating the boundary of sufficiency—nodes to the left of the line are insufficient. Nodes represent one or several sub-observations. Solid arrows indicate known child relationship between the elements of the two nodes, whereas dotted arrows indicate potential children if the right members in each node are selected. The rounded rectangle node represents the critical observations.

4.2. A generic algorithm

With this scaffolding erected, we are now positioned to define a procedure to compute critical observations. This first procedure is still very simple and improvements are proposed in Section 6. We assume that there exists a procedureMinDiagnose that is able to compute the diagnosis $\Delta(\theta)$ of a sub-observation θ for the particular modelling framework.

Under this assumption and Assumption 2, Procedure findCriticalObservation, presented in Fig. 4, is guaranteed to return a critical observation.

The algorithm proceeds by computing the children of the initial sub-observation (by construction, this initial $\theta = sub(\mathcal{O})$ will have the same diagnosis), and progressively evaluates the induced sub-problems. Should a sufficient child be found, all other candidates at the current level of abstraction are discarded and new children are generated from this point. In this way the algorithm is greedy in nature, as we immediately consider any newly sufficient sub-observation the best current candidate we are liable to miss potentially more concise sub-observations on different chains.

Theorem 4.3. *Given a diagnosis problem $P = \langle \mathcal{D}, \mathcal{O} \rangle$, then* findCriticalObservation *is guaranteed to find a critical observation, assuming that* MinDiagnose$(\cdot)$ *is terminating and deterministic, $children(\cdot)$ is defined, and Assumption 2 holds. In other words,* findCriticalObservation *is correct, complete, and terminating.*

The algorithm can be modified to find all critical observations by maintaining the full list of candidates as opposed to greedily keeping only those most recently found, and remembering any found critical observations. The algorithm is intentionally generic to allow for further extension and refinement, and as such doesn't take into account setting-specific optimisations. As a result, the algo-

```
Procedure findCriticalObservation
input: $obs$ := observation
output: critical observation
$diag$ := MinDiagnose($obs$)
$\theta$ := $sub(obs)$
$candidates$ := $children(\theta)$
while $candidates \neq \emptyset$ do
    $\theta'$ := remove_one($candidates$)
    if $diag$ = MinDiagnose($\theta'$) then
        $\theta$ := $\theta'$
        $candidates$ := $children(\theta)$
    end if
end while
return $\theta$
```

Fig. 4. A generic algorithm for finding a critical observation for a diagnosis sub-problem. The model is implicit.

rithm is likely going to be sub-optimal in most settings. One possible improvement is the trimming of the search space heuristically when insufficient and critical sub-observations are found, but such a modification can substantially increase the complexity if not carefully considered.

Example. We execute findCriticalObservation on the running example, setting:

$$\mathcal{O} = \left\{s_0 \mapsto N, s_1 \mapsto N, s_2 \mapsto \boldsymbol{F}, s_3 \mapsto \boldsymbol{F}, s_4 \mapsto N \ldots, s_7 \mapsto N\right\}$$

The corresponding minimal diagnosis is $\Delta_{\min}(\mathcal{O}) = \left\{\left\{x_2, x_3\right\}, \left\{b_3\right\}\right\}$. We recall that the initial sub-observation is

$$\theta = \left\{\langle s_0, N\rangle, \langle s_1, N\rangle, \langle s_2, \boldsymbol{F}\rangle, \langle s_3, \boldsymbol{F}\rangle, \langle s_4, N\rangle, \langle s_5, N\rangle, \langle s_6, N\rangle, \langle s_7, N\rangle\right\}.$$

The set of children of θ are computed, and they correspond to the second row of Fig. 3, i.e., $\theta_{[0]}, \ldots, \theta_{[7]}$. We will consider the case that the algorithm sorts these in lexical order, and therefore $\theta_{[0]}$ is considered first. The minimal diagnosis of $\theta_{[0]}$ equals $\Delta_{\min}(\mathcal{O})$; therefore $\theta_{[0]}$ is sufficient and it replaces θ as the current root of the search tree in the algorithm. The set of children of $\theta_{[0]}$ is updated (here, $\theta_{[0,1]}, \ldots, \theta_{[0,7]}$).

Per lexical ordering, $\theta_{[0,1]}$ is selected next for evaluation and diagnosing: $\Delta_{\min}(\theta_{[0,1]}) = \left\{\left\{x_2, x_3\right\}, \left\{b_3\right\}, \left\{b_0\right\}\right\}$. Since it differs from $\Delta_{\min}(\mathcal{O})$, it is ignored and $\theta_{[0,2]}$ is considered instead. This sub-observation, as well as $\theta_{[0,3]}$, both turn out to be insufficient, but $\theta_{[0,4]}$ is sufficient and replaces $\theta_{[0]}$. The children of this sub-observation are:

$$\left\{\theta_{[0,1,4]}, \theta_{[0,2,4]}, \theta_{[0,3,4]}, \theta_{[0,4,5]}, \theta_{[0,4,6]}, \theta_{[0,4,7]}\right\}.$$

In the next step, $\theta_{[0,1,4]}$ is considered first. Notice that it is "trivially" insufficient since it is a child of $\theta_{[0,1]}$ which was already proved to be insufficient. Identifying such relations is not always obvious however, and this optimisation is discussed in Section 6. Eventually, $\theta_{[0,4,5]}$ is shown to be sufficient.

The algorithm runs until it finds the sufficient sub-observation $\theta_{[0,4,5,6,7]}$. The children of this sub-observation are $\left\{\theta_{[0,1,4,5,6,7]}, \theta_{[0,2,4,5,6,7]}, \theta_{[0,3,4,5,6,7]}\right\}$, all of which are insufficient. Therefore, $\theta_{[0,4,5,6,7]}$ is critical, and findCriticalObservation terminates and returns this result.

4.3. Framework instantiation requirements

The framework, as codified here, is intentionally generic in nature so that it can be applied to diagnosis problems using different modelling techniques. To instantiate the framework for a given setting and problem, the following is required:

- A clear idea of what a sub-observation in a given scenario looks like.
- This gives rise to the space of sub-observations, Θ, the collection of all possible sub-observations.
- This is then equipped with a partial ordering relationship, $\preceq$, over the sub-observations in Θ
- And further equipped with an injective function $sub(\cdot)$ that maps observations produced by the system in question to a corresponding sub-observation $\theta \in \Theta$.

From there, a few additional criteria need to be satisfied before a Critical Observation can be found algorithmically:

- The ordering relation should guarantee that the set of sub-observations that are preferred to any given θ is finite.
- There is a defined diagnosis procedure, MinDiagnose, that behaves predictably on both sub-observations and observations.
- We have a well-defined and well-ordered familial function, $children(\cdot)$, that produces all of the sub-observations that are exactly one additional level of abstraction away from a given θ. Such a function will necessarily determine what one level of abstraction away looks like.

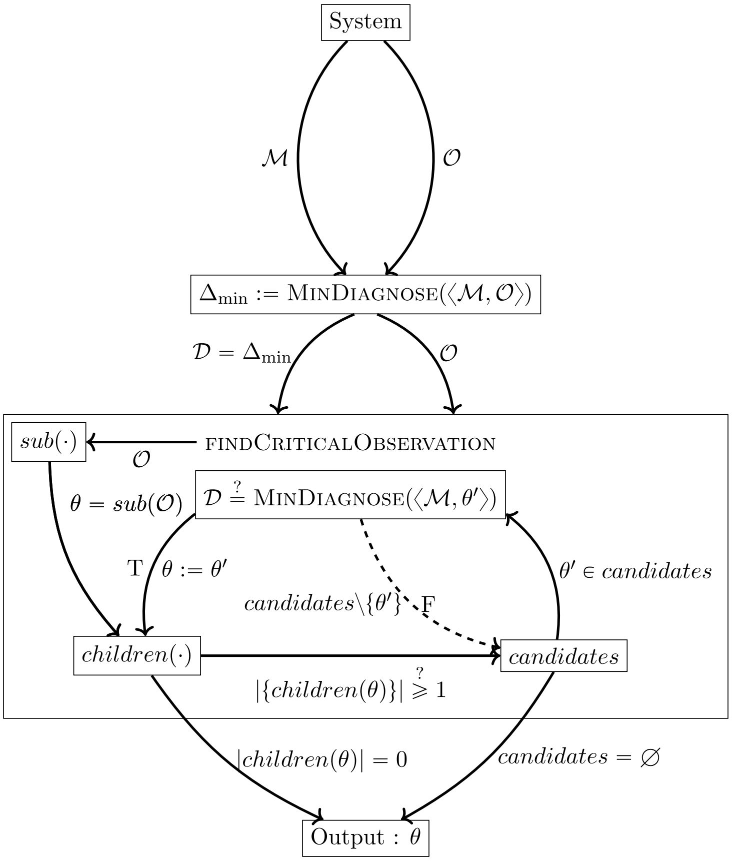
System
$\mathcal{M}$
$\mathcal{O}$
$\Delta_{\min} := \textsc{MinDiagnose}(\langle \mathcal{M}, \mathcal{O} \rangle)$
$\mathcal{D} = \Delta_{\min}$
$\mathcal{O}$
$sub(\cdot)$
$\mathcal{O}$
findCriticalObservation
$\theta = sub(\mathcal{O})$
$\mathcal{D} \overset{?}{=} \textsc{MinDiagnose}(\langle \mathcal{M}, \theta' \rangle)$
T
$\theta := \theta'$
$\theta' \in candidates$
$candidates \backslash \{\theta'\}$
F
$children(\cdot)$
$candidates$
$|\{children(\theta)\}| \overset{?}{\geqslant} 1$
$|children(\theta)| = 0$
$candidates = \varnothing$
Output : θ

Fig. 5. The Framework.

We summarise the framework and algorithm in Fig. 5, showing the flow of input and output from the initial system specification through to the generation of critical observations. Inside the box representing the algorithm, the edge labels indicate assignments or conditional checks as appropriate.

With the framework instantiated according to these requirements, we are guaranteed to find a critical observation. However, the generic algorithm is inefficient by nature, as it seeks to accommodate all model-based paradigms. For specific settings, the algorithm can be optimised and extended to take advantage of properties inherent to that setting, some of which are described in subsequent sections of this work.

5. Instantiating the framework

In this section we show how the theory of critical observations and the associated framework can be used to derive diagnostic explanations in the context of state-based observations and event-based observations.

5.1. State-based diagnosis

5.1.1. Instantiation

We make reference to the first principles framework for model-based diagnosis presented in Section 2 (specifically the original consistency-based work of [1]), and we take the definition of the diagnoser from this unmodified.

An observation is a function $\mathcal{O} : \mathcal{V} \to D$ where $\mathcal{V}$ is a finite set of variables and D is the domain of the variables. For dynamic systems, a variable $v \in \mathcal{V}$ could be timed (e.g., the temperature at a given time).

We define the required aspects of the framework in Definition 6:

Definition 12 *(State-based instantiation)*. The state-based instantiation $\langle \Theta_s, \preceq_s, sub_s \rangle$ of the sub-observation framework as proposed in Definition 6 is:

- Θ_s is the set of finite sets θ of *assignments* where an assignment is a pair $\langle v, d \rangle$ in which v is a *variable* and d is the *value* assigned to this variable;

- $\preceq_s$ is the traditional subset membership: $\preceq_s \overset{def}{=} \subseteq$;
- sub_s is the function that returns all the assignments in the observation function: $sub_s(\mathcal{O}) = \{\langle v, d\rangle \mid \mathcal{O}(v) = d\}$.

This instantiation of the framework is one of the more intuitive as it deals with standard mathematical objects we are very familiar with.[3] The remaining requirements for use of the generic algorithm follow by virtue of common, known, results:

- $\langle \Theta_s, \preceq_s\rangle$ is finite.
- The MinDiagnose algorithm can be implemented by considering the sub-observation θ as an observation over a subset of observable variables.
- A sub-observation θ' is a child of another sub-observation θ if and only if it is a subset of the latter with cardinality of exactly one less. In other words, the set of children can be computed by:

$$children(\theta) = \{\theta \setminus \{\langle v, d\rangle\} \mid \langle v, d\rangle \in \theta\}\,.$$

Other possible abstractions could be defined. For instance, for numerical observations, an exact value could be replaced by a range or a qualitative value. For instance, in the context of a heating system the observed value $\left\langle t_{pipe}, 2.1°\mathrm{C}\right\rangle$ could be replaced with an abstract fragment like “pipe is cold”.

5.1.2. Computing a first sufficient sub-observation

We now show an optimisation specifically targeted to state-based observations and a diagnosis framework in which the hypothesis space is a powerset of faults, that allows us to start from some θ that has already been partially abstracted, instead of the usual starting point given by $sub(\cdot)$.

As is classically the case [1,2], we assume that a hypothesis is a subset of faults (the faults that are currently affecting the system), and that a minimal candidate should include a (subset-)minimal set of faults:

- $\mathbb{H} = 2^F$ for some set F of faults and
- $h_1 \preceq_{\mathbb{H}} h_2$ iff $h_1 \subseteq h_2$.

A *conflict* is a subset $C \subseteq F$ of faults that intersects all diagnosis candidates ($\forall \delta \in \Delta.\ C \cap \delta \neq \emptyset$). It is known that the set of minimal candidates is precisely the set of hitting sets of the minimal candidates [1,2].

We introduce the notion of “observable conflict” that decorates a conflict with a set of observed assignments:

Definition 13 *(Observable conflict).* Let $\mathcal{P} = \langle \mathcal{D}, \mathcal{O}\rangle$ be a diagnosis problem where $\mathcal{D}$ is a diagnosis framework over the set F of faults. An *observable conflict* is a pair $\mathcal{X} = \langle C, \theta\rangle$ where $\theta \subseteq \mathcal{O}$ is a set of assignments such that C is a conflict for any observation $\mathcal{O}'$ that satisfies $\theta \subseteq \mathcal{O}'$.

The following theorem allows us to compute a sufficient sub-observation.

Theorem 5.1. *Let $\mathcal{P} = \langle \mathcal{D}, \mathcal{O}\rangle$ be a diagnosis problem where $\mathcal{D}$ is a diagnosis framework over the set F of faults. Let $\{\mathcal{X}_1, \dots, \mathcal{X}_k\}$ be a set of observable conflicts of $\mathcal{P}$ with, for all i, $\mathcal{X}_i = \langle C_i, \theta_i\rangle$ such that $\{C_1, \dots, C_k\}$ is the set of minimal conflicts of $\mathcal{P}$. Then, θ is a sufficient sub-observation for $\mathcal{P}$ where*

$$\theta = \bigcup_{i \in \{1,\dots,k\}} \theta_i.$$

Example. Consider our running example from Fig. 1. This example features two minimal conflicts: $C_1 = \left\{b_3, x_2\right\}$ and $C_2 = \left\{b_3, x_3\right\}$. A (minimal) support for C_1, i.e., a sub-observation θ_1 such that $\langle C_1, \theta_1\rangle$ forms an observable conflict is $\theta_1 = \left\{\langle s_0, N\rangle, \langle s_2, \boldsymbol{F}\rangle\right\}$; another minimal support is $\theta_1' = \left\{\langle s_1, N\rangle, \langle s_2, \boldsymbol{F}\rangle\right\}$. Similarly, both $\theta_2 = \left\{\langle s_0, N\rangle, \langle s_3, \boldsymbol{F}\rangle\right\}$ and $\theta_2' = \left\{\langle s_1, N\rangle, \langle s_3, \boldsymbol{F}\rangle\right\}$ entail C_2.

We can then assume here that we end up with the set of observable conflicts $\left\{\langle C_1, \theta_1\rangle, \left\langle C_2, \theta_2'\right\rangle\right\}$. From Theorem 5.1, we have $\theta \overset{def}{=} \theta_1 \cup \theta_2' = \left\{\langle s_0, N\rangle, \langle s_1, N\rangle, \langle s_2, \boldsymbol{F}\rangle, \langle s_3, \boldsymbol{F}\rangle\right\}$ is a sufficient sub-observation. We note however that θ, while sufficient, is not critical. Still, θ is already substantially abstracted from the original observation $\mathcal{O}$, which simplifies the task of computing a critical sub-observation.

The possibility of generalising this optimisation it to a larger class of observations, whilst interesting and possible, is outside the scope of this work.

[3] The state-based framework was indeed the catalyst for the naming convention of sub-observations, as they are given by subsets in this setting.

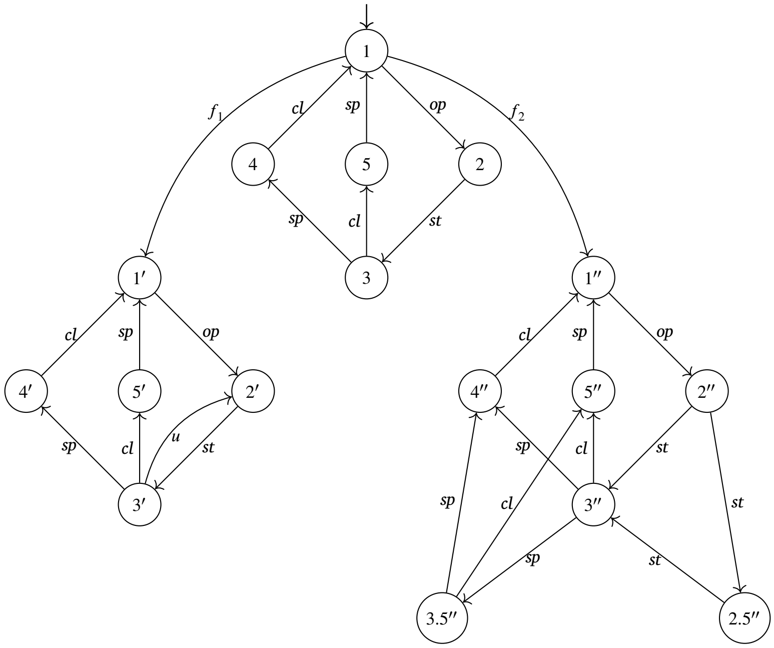

Fig. 6. Example of fan-vent system.

5.2. Event-based diagnosis

We now move to event-based diagnosis, in particular applied to discrete event systems (DES). In Subsection 5.2.1, we provide an example and show that it contains *implicit* information which additional care must be taken to handle. Then in Subsection 5.2.2, we provide an event-based sub-observation framework based on the notion of hard and soft events. In Subsection 5.2.3, we show that the intuitive definition of this framework is correct, which allows us to propose a diagnosis algorithm for a sub-observation. Finally in Subsection 5.2.4, we show how to compute the children of a sub-observation, an essential component to find critical observations.

5.2.1. Example

We provide here an event-based example that we use throughout this section.

Consider the system modelled in Fig. 6, which is inspired by the HVAC system from [10]. This model is an untimed discrete event system (DES, [11]). Circles represent states, and edges represent state changes; for instance, the edges $1 \xrightarrow{op} 2$ indicate that the occurrence of the *op* event in state 1 leads the system to state 2. A behaviour of the system is a sequence of transitions denoted $s_0 \xrightarrow{e_1} s_1 \xrightarrow{e_2} \ldots \xrightarrow{e_k} s_k$.

During a normal cycle (represented by the states 1 to 5), the controller opens (*op*) the valve, starts (*st*) the pump, stops (*sp*) it, and then closes (*cl*) the valve (although the last two commands can be reordered). In the first faulty condition (f_1), the pump sometimes shuts (*u*) itself down. In the second faulty condition, the valve jams and up to two attempts are necessary to open or close it.

Given behaviour $w = s_0 \xrightarrow{e_1} \ldots \xrightarrow{e_k} s_k$, the associated observation $\mathfrak{F}_{\mathbb{O}}(w)$ is the sequence of events projected onto the *observable events*, here $\{op, st, sp, cl\}$.

We consider the following four behaviours:

(a) $w_1 = 1 \xrightarrow{op} 2 \xrightarrow{st} 3 \xrightarrow{sp} 4 \xrightarrow{cl} 1$ (nominal):
- observation $\mathcal{O}_1 = op\ st\ sp\ cl$;
- diagnosis $\Delta_1 = \{\emptyset, \{f_1\}, \{f_2\}\}$;
- minimal diagnosis $\Delta_{\min 1} = \{\emptyset\}$.

(b) $w_2 = 1 \xrightarrow{f_1} 1' \xrightarrow{op} 2' \xrightarrow{st} 3' \xrightarrow{u} 2' \xrightarrow{st} 3' \xrightarrow{sp} 4' \xrightarrow{cl} 1'$:
- observation $\mathcal{O}_2 = op\ st\ st\ sp\ cl$;
- diagnosis $\Delta_2 = \{\{f_1\}, \{f_2\}\}$;
- minimal diagnosis $\Delta_{\min 2} = \{\{f_1\}, \{f_2\}\}$.

(c) $w_3 = 1 \xrightarrow{f_1} 1' \xrightarrow{op} 2' \xrightarrow{st} 3' \xrightarrow{u} 2' \xrightarrow{st} 3' \xrightarrow{u} 2' \xrightarrow{st} 3' \xrightarrow{sp} 4' \xrightarrow{cl} 1'$:
- observation $\mathcal{O}_3 = op\ st\ st\ st\ sp\ cl$;
- diagnosis $\Delta_3 = \{\{f_1\}\}$;
- minimal diagnosis $\Delta_{\min 3} = \{\{f_1\}\}$.

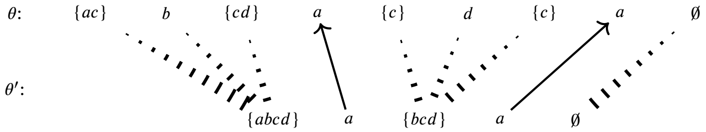

Fig. 7. An example mapping two sub-observations satisfying $\preceq$.

(d) $w_4 = 1 \xrightarrow{f_2} 1'' \xrightarrow{op} 2'' \xrightarrow{st} 3'' \xrightarrow{sp} 3.5'' \xrightarrow{sp} 4'' \xrightarrow{cl} 1''$:
- observation $\mathcal{O}_4 = op\ st\ sp\ sp\ cl$;
- diagnosis $\Delta_4 = \{\{f_2\}\}$;
- minimal diagnosis $\Delta_{\min 4} = \{\{f_2\}\}$.

Looking at the observations returned in each behaviour, we note that, in contrast with their state-based counterpart, the information contained in the event-based observations is not so much in what happened (all four observable events were observed in all four scenarios) but how often they happened and what happened between them. For instance, in the third scenario, three *st* occurring in succession (without any other event between them) is symptomatic of fault f_1.

5.2.2. Framework instantiation

We assume that the observation space is the Kleene closure of a set of *observable events* Σ_o, so that an observation is a sequence of observable events: $\mathcal{O} \in \Sigma_o^*$.

Our definition of sub-observation is given in Definition 15, but we give an intuitive definition first. The definition relies on the notion of *hard* (or firm), and *soft* (or possible) events. In a sub-observation, a hard event is an event that did occur; a soft event is a set of events that may have occurred any number of times, with the idea that assuming the occurrence of these events does not change the (minimal) diagnosis.

Definition 14 *(Hard and soft events).* A *hard event* is a single observable event $x \in \Sigma_o$. A *soft event* is a (possibly empty) subset of observable events $y \subseteq \Sigma_o$.

With this definition, we are now in a position to define (Definition 15) a sub-observation as a sequence of intertwined soft and hard events $\theta = y_0 x_1 \ldots x_k y_k$. From the perspective of the observed sequence, the events $x_1, \ldots, x_k$ should appear in the sequence in this order, and the events between x_i and x_{i+1} should be captured in the soft event y_i. From the perspective of the sub-observation, it is known that some (possibly none) events from y_0 took place, followed by x_1, followed again by some events from y_1, etc. Intuitively, and in other words, the sub-observation $\theta = y_0 x_1 \ldots x_k y_k$ represents all the observations captured by the regular expression $(y_0^*) x_1 \ldots x_k (y_k^*)$.

The precedence relation $\preceq$ and the injection function *sub* follow logically from the semantics. Given observation $\mathcal{O} = o_1 \ldots o_k$, the injection function $sub(\mathcal{O})$ ought to return the least abstract sub-observation which includes $\mathcal{O}$, i.e., the sub-observation that indicates exactly what was observed. This is $sub(\mathcal{O}) = y_0 x_1 \ldots x_k y_k$ where each x_i equals o_i (these events were definitely observed) and each y_i is the empty set (we do not include any uncertainty).

The precedence relation $\preceq$ is more technical in this setting, but still fairly simple to explain. Given two sub-observations $\theta = y_0 x_1 \ldots x_k y_k$ and $\theta' = y'_0 x'_1 \ldots x'_{k'} y'_{k'}$, then for $\theta' \preceq \theta$ to hold, we require:

- all x'_i appear in some x_j, and they appear in the same order (in Definition 15, this is guaranteed by the strict monotonicity of the function f),
- the hard events x_j that have no x'_i counter-part should be captured in some y'_i, and
- the soft event y_j should also be captured in some y'_i.

Fig. 7 illustrates this definition, showing that $\theta' = \{a, b, c, d\}\, a\, \{b, c, d\}\, a \emptyset$ is an abstraction of $\theta = \{a, c\}\, b\, \{c, d\}\, a\, \{c\}\, d\, \{c\}\, a \emptyset$. Indeed all sequences of observations that θ allows are also allowed by θ'. We can notice that θ' allows for more sequences than θ however, as, for instance, it does not enforce the occurrence of b, it allows d to be the first event, and it allows the occurrence of b between the two subsequent occurrences of a.

Definition 15 *(Event-based instantiation).* The framework $\langle \Theta, \preceq, sub \rangle$ for observed sequences $\mathcal{O} \in \Sigma_o^*$ is defined by:

- $\Theta = 2^{\Sigma_o} \left(\Sigma_o 2^{\Sigma_o} \right)^*$ such that each sub-observation is a sequence $\theta = y_0 x_1 y_1 x_2 \ldots x_k y_k$ with $x_i \in \Sigma_o$ and $y_j \subseteq \Sigma_o$; the *length* of θ is $|\theta| = k$;
- $\theta' \preceq \theta$ iff, given $|\theta'| = n$, $|\theta| = m$, there exists a mapping function $f : \{0, \ldots, n+1\} \mapsto \{0, \ldots, m+1\}$ satisfying:
  - $\forall i \in \{0, \ldots, n\} .\ f(i) < f(i+1)$
  - $f(0) = 0$
  - $f(n+1) = m+1$
  - $\forall i \in \{0, \ldots, n\} .\ x'_i = x_{f(i)}$

  - $\forall i \in \{0, \ldots, n\} \,.\, y'_i \supseteq \bigcup_{f(i) \leq j \leq f(i+1)-1} y_j \cup \bigcup_{f(i) < j < f(i+1)} x_j$;
- if $\mathcal{O} = o_1 \ldots o_k$ then $sub(\mathcal{O}) = \emptyset o_1 \emptyset \ldots \emptyset o_k \emptyset$.

We illustrate these definitions on the example of Fig. 6.

The first observation is $\mathcal{O}_1 = op\ st\ sp\ cl$, with minimal diagnosis $\Delta_{\min 1} = \{\emptyset\}$. As is generally the case when the minimal diagnosis contains the most preferred candidate, the critical sub-observation for this problem is the most abstract one: $\theta_0 = \Sigma_{\text{o}}$, which specifies that an unspecified number (which includes zero) of observable events occurred.

The second observation is $\mathcal{O}_2 = op\ st\ st\ sp\ cl$ with minimal diagnosis $\Delta_{\min 2} = \{\{f_1\}, \{f_2\}\}$. In this case, one of the critical sub-observations is $\theta_2 = \Sigma_{\text{o}}\ st\ \{op, st, cl\}\ st\ \Sigma_{\text{o}}$, which specifies that the command *st* was issued twice without the issuance of a *sp* in between. This demonstrates that the system cannot be in nominal behaviour as one expects at least one *stop* command between two *starts*. Another possible critical sub-observation is $\theta'_2 = \Sigma_{\text{o}}\ st\ \{st, sp, cl\}\ st\ \Sigma_{\text{o}}$.

The third observation is $\mathcal{O}_3 = op\ st\ st\ st\ sp\ cl$ with minimal diagnosis $\Delta_{\min 3} = \{\{f_1\}\}$. This time, one of the critical sub-observations is $\theta_2 = \Sigma_{\text{o}}\ st\ \{op, st, cl\}\ st\ \{op, st, cl\}\ st\ \Sigma_{\text{o}}$, which indicates that there were three *st* without a single *sp*. Compared to the previous example, we now rule out f_2 which only allows two consecutive *stop* commands.

The fourth observation is $\mathcal{O}_4 = op\ st\ sp\ sp\ cl$ with minimal diagnosis $\Delta_4 = \{\{f_2\}\}$. One critical sub-observations is $\theta_4 = \Sigma_{\text{o}}\ sp\ \{st, sp, cl\}\ sp\ \Sigma_{\text{o}}$, which indicates that there were two occurrences of *sp* without an occurrence of *op*.

5.2.3. Formalising the intuition

We presented the intuition that $\theta = y_0 x_1 \ldots x_k y_k$ represents all observations that include the hard events x_i in this order, with soft events y_i between x_i and x_{i+1}. We now formally establish this result.

Theorem 5.2. *Let $\theta = y_0 x_1 \ldots x_k y_k$ be a sub-observation. The set $\psi(\theta)$ is defined by the regular expression $y_0^* x_1 \ldots x_k y_k^*$.*

From this intuition, we can define an algorithm for computing the diagnosis of a sub-observation. The literature on diagnosis of DES is rich since the seminal work of Sampath et al. [10]. Traditionally, the model has been described as an automaton, and diagnosis is performed by following the trajectories on the automaton that match the observation [12]. Different techniques have been proposed to alleviate the computational cost, such as decentralised/distributed [13–15], the use of Petri nets [16], SAT [17], AI planning [18]. See [19] for a review of existing approaches. Importantly, it is possible to view sub-observations as "uncertain observations" which can be represented as an automaton [20] and handled by most diagnosers.

5.2.4. Traversal

With the necessary results in place, we now require a method by which we can determine the children of a sub-observation, that subsequently allows us to traverse the search space. To this end, we introduce two new operations and prove that the sub-observations obtained from these operations are all those that are minimally more-abstracted with respect to the given sub-observation. That is to say, there are no potential sub-observations that can be between a parent and the generated children, and there are no additional children beyond those produced by these operations.

The operations are respectively called *event-softening*, and *collapse*, which we define:

Definition 16 *(Event softening)*. Given sub-observation $\theta = y_0 x_1 \ldots x_k y_k$, the *event-softening* operation $\theta' = es(\theta, i, e)$ adds event e to the ith soft event of the sub-observation: $es(\theta, i, e) = y'_0 x'_1 \ldots x'_k y'_k$ (defined if $e \notin y_i$) such that:

- $\forall j \in \{1, \ldots, k\} \,.\, x'_j = x_j$,
- $\forall j \in \{0, \ldots, k\} \setminus \{i\} \,.\, y'_j = y_j$, and
- $y'_i = y_i \cup \{e\}$.

We illustrate event-softening with a simple example:

Take $\theta = \{abcd\}\, a\, \{bcd\}\, a \emptyset$. Suppose we wish to soften the middle of the behaviour as we anticipate that a might perhaps be irrelevant. We would then call $es(\theta, 1, a)$, producing $\theta' = \{abcd\}\, a\, \{abcd\}\, a \emptyset$.

Definition 17 *(Collapse)*. Given a sub-observation $\theta = y_0 x_1 \ldots x_k y_k$, the *collapse* operation $\theta' = coll(\theta, i)$ "forgets" the concrete occurrence of a hard event x_i. This operation requires the soft events before and after x_i to be equal and to allow for x_i: $coll(\theta, i) = y'_0 x'_1 y'_1 \ldots x'_{k-1} y'_{k-1}$ (defined if $x_i \in y_i$ and $y_{i-1} = y_i$) such that:

- $\forall j \in \{1, \ldots, i-1\} \,.\, x'_j = x_j$ and $y'_{j-1} = y_{j-1}$
- $\forall j \in \{i+1, \ldots, k\} \,.\, x'_{j-1} = x_j$ and $y'_{j-1} = y_j$, and
- $y'_{i-1} = y_{i-1} = y_i$.

Continuing the example used to illustrate event-softening, we now show collapse on a hard event in θ'. A requirement for collapse as defined is that the soft events on either side of the hard event selected for collapse contain said hard event. We see this in θ', as

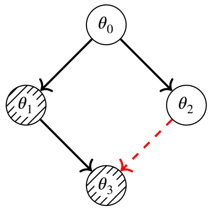

Fig. 8. Fragment of $\Theta(\mathcal{O})$ illustrating how some sub-observations can be ignored: because θ_3 is a sub-observation of θ_1 and θ_1 is known to be not sufficient, θ_3 can be automatically labelled as not sufficient.

```
Procedure findCriticalObservation
input: $diag$ := minimal diagnosis
input: $\theta$ := sub-observation defined as a set of assignments
output: critical observation
for all $o \in \theta$ do
    if $diag$ = MinDiagnose($\theta \setminus \{o\}$) then
        $\theta := \theta \setminus \{o\}$
    end if
end for
return $\theta$
```

Fig. 9. Procedure for finding a critical observation when sub-observations are sets of assignments and the preference is the subset relation.

both y_0 and y_1 contain a. The second condition is that both of the relevant soft events are equal, and in this case both are $\{abcd\}$. The result of $coll(\theta', 1)$, then, is $\theta'' = \{abcd\}\, a\emptyset$.

The following lemma states that these two operations together are complete:

Lemma 5.3. *The children of a sub-observation θ are exactly all the sub-observations that can be obtained by applying either event-softening or collapse to θ.*

6. Algorithmic optimisation

We now discuss implementations issues. We propose two improvements to the procedure of computing the critical observation and then broadly discuss complexity results pertaining to the algorithm.

6.1. Pruning the search space

One possible enhancement relies on the structure of the sub-observation space to identify that some of the children of the current sufficient sub-observation can be ignored. We base our explanation on the example of Fig. 8. In this figure, the arrows indicate a child/parent relation between the sub-observations, and struck-out circles indicate sub-observations that are not sufficient. Assume that θ_0 is a sub-observation known to be sufficient but whose criticality status is unknown. Child θ_1 of θ_0 is then first verified and proved to be not sufficient. Next, child θ_2 is verified and shown to be sufficient. From the monotonicity property and the fact that θ_3 is a child of θ_1, we deduce that θ_3 will also not be sufficient. Consequently, it is possible to automatically ignore it.

Depending on the definition of the sub-observation space, the pruning rules will differ. For instance, in the framework of state-based observations where a sub-observation is a subset of atomic observations and the abstraction relation is the subset relation, we have the following result:

Lemma 6.1. *Let $\mathcal{P} = \langle \mathcal{D}, \mathcal{O} \rangle$ be a diagnosis problem and let $\mathcal{S}_s$ be the state-based instantiation of the sub-observation framework from Definition 12. Let θ be a sub-observation of $\mathcal{O}$ in the framework $\mathcal{S}_s$. Let $o \in \theta$ be an assignment. If $\theta \setminus \{o\}$ is not sufficient, then for any sub-observation θ' of θ (i.e., such that $\theta' \subseteq \theta$), $\theta' \setminus \{o\}$ is not sufficient.*

An interpretation of Lemma 6.1 is that once an atomic observation has been proved to be necessary in a sub-observation, it will necessarily be in any critical observation that is derived from this sub-observation. It is therefore possible to use the algorithm presented in Fig. 9 in which each atomic observation is tested exactly once. This implies that the number of calls to MinDiagnose drops from $O(n^2)$ to $O(n)$ where n is the cardinality of θ.

Similar results can be derived for event-based sub-observations as we defined them in Subsection 5.2, but they are more complicated. They are summarised in the following lemma, for which we first define new notations:

- Given a sub-observation θ, we write shorthand $es_{i,e}(\theta) \overset{def}{=} es(\theta, i, e)$ and $coll_i(\theta) \overset{def}{=} coll(\theta, i)$ such that it is possible to chain these operators. For instance, $es(es(\theta, i, e), i', e')$ will be equivalently written $es_{i,e} \circ es_{i',e'}\ (\theta)$.
- We also propose a "tracking" operator tr_{op} where op is an operator $es_{i,e}$ or $coll_i$ which takes as input an index j and returns an index defined by:

- $tr_{es_{i,e}}(j) \overset{def}{=} j$ for all i, e;
- $tr_{coll_i}(j) \overset{def}{=} j - 1$ if $i < j$;
- $tr_{coll_i}(j) \overset{def}{=} j$ if $i \geq j$.

The tracking operator allows us to track a position in the sub-observation after changes are performed. For instance, an event softening $es_{i,e}$ followed by a collapse $coll_{i'}$ applied to θ, is equivalent to a collapse $coll_{i'}$ followed by an event softening $es_{j,e}$ where $j = tr_{coll_{i'}}(i)$ (assuming these operations are allowed).

Lemma 6.2. *Let θ be an event-based sub-observation as defined in Subsection 5.2. Let $\theta' = op_1 \circ \ldots \circ op_k(\theta)$ be an abstraction of θ. Let i be an index and $j = tr_{op_1} \circ \ldots \circ tr_{op_k}(i)$.*

- *If $es(\theta, i, e)$ is not sufficient, then $es(\theta', j, e)$ is not sufficient.*
- *If $coll(\theta, i)$ is not sufficient, then $coll(\theta, j)$ is not sufficient.*

Consider for instance the example of Fig. 6 along with the observation $\mathcal{O}_4 = op\ st\ sp\ sp\ cl$, and sub-observation $\theta_0 = \Sigma_o\ op\ \Sigma_o\ sp\ \{op, cl\}\ sp\ \Sigma_o$, which specifies that a) there was at least one op and b) there later was two sp without an st in between; only faulty event f_2 can explain this. Sub-observation $\theta_1 \overset{def}{=} es(\theta_0, 2, st) = \Sigma_o\ op\ \Sigma_o\ sp\ \{op, cl, st\}\ sp\ \Sigma_o$ is not sufficient, because it removes the decisive piece of information (b). Instead, sub-observation $\theta_2 \overset{def}{=} coll(\theta_0, 1) = \Sigma_o\ sp\ \{op, cl\}\ sp\ \Sigma_o$ is sufficient as it removes the irrelevant piece of information (a). The tracking of index 2 gives us an index of 1 or, formally, $tr_{coll_1}(2) = 1$. In other words, we do not need to prove that $\theta_3 \overset{def}{=} es(\theta_2, 1, st) = \Sigma_o\ sp\ \{op, cl, st\}\ sp\ \Sigma_o$, by construction known to be an abstraction of θ_1, is not sufficient.

Given n observed events and a set Σ_o of m observable events, it is possible to reduce the number of calls to MinDiagnose to $O(nm)$.

6.2. Not computing the diagnosis

The last enhancement that we propose applies when the sufficiency of a sub-observation needs to be verified. We first recall that θ is sufficient if and only if $\Delta_{\min}(\theta)$ equals $\Delta_{\min}(\mathcal{O})$. We show however that it is not always necessary to compute the minimal diagnosis associated with the current sub-observation in order to prove or disprove sufficiency.

Example. We first illustrate this notion with an example using set-inclusion minimality to determine candidate hypotheses. We assume that the diagnosis problem is defined on a classical diagnosis framework over the set $F = \{a, b, c, d\}$ of faults.

For some execution of the system we find that $\Delta_{\min} = \{\{d\}, \{a, b\}\}$, i.e., either d is faulty or both a and b. Fig. 10 shows the hypothesis space partitioned into four groups depending how each hypothesis compares to the minimal candidates: the minimal candidates, the hypotheses that strictly precede through the $\subseteq$ relation some minimal candidate; the hypotheses that are strictly preceded by some minimal candidate; and the other hypotheses (that are not related to any minimal candidate through $\subseteq$).

Assume now that we want to test whether sub-observation θ is sufficient. By Definition 9, θ is sufficient if and only if $\Delta_{\min}(\theta)$ equals $\Delta_{\min}(\mathcal{O})$. From the monotonicity property of the diagnosis with respect to sub-observations, we know that all candidates of $\mathcal{O}$ are also candidates of θ. Therefore, in order for $\Delta_{\min}(\theta)$ to differ from $\Delta_{\min}(\mathcal{O})$, one of the hypotheses in orange or red in Fig. 10 ought to be a candidate.

The result of this example is formalised through the following theorem.

Theorem 6.3. *Let $\mathcal{P} = \langle \mathcal{D}, \mathcal{O} \rangle$ be a diagnosis problem such that $\mathcal{D} = \langle \mathcal{M}, \mathbb{O}, \mathbb{H}, \mathfrak{F}_{\mathbb{O}}, \mathfrak{F}_{\mathbb{H}}, \preceq_{\mathbb{H}} \rangle$. Let θ be a sub-observation of $\mathcal{O}$. Then $\Delta_{\min}(\theta) = \Delta_{\min}(\mathcal{P})$ holds iff $\Delta(\theta) \cap \nabla = \emptyset$ holds for*

$$\nabla = \left\{ h \in \mathbb{H} \mid \forall \delta \in \Delta_{\min}(\mathcal{P}).\ \delta \npreceq_{\mathbb{H}} h \right\}.$$

As our goal here is to verify whether any of the hypotheses in ∇ is a candidate, we attempt to leverage the result from Theorem 6.3. This set, however, can be very large in general, and as such testing each of these hypotheses individually is not recommended. We give an example that shows how this can be implemented efficiently in practice.

Example. Continuing the previous example of Fig. 10, we recall that the minimal diagnosis is $\Delta_{\min} = \{\{d\}, \{a, b\}\}$. This implies that ∇ is $\left\{\emptyset, \{a\}, \{b\}, \{c\}, \{a, c\}, \{b, c\}\right\}$. This set is precisely all subsets of $\{a, b, c, d\}$ that do not contain d and that do not contain both a and b. If diagnosis is performed by using a logic solver, then $\Delta(\theta) \cap \nabla$ is empty if and only if the two following statements are both inconsistent with the model and the sub-observation: $\neg d$ and $\neg(a \wedge b)$.

In other words, the minimal conflicts of θ should be the same as the minimal conflicts of $\mathcal{P}$, and it is sufficient to verify that the minimal conflicts of $\mathcal{P}$ are also minimal conflicts of θ.

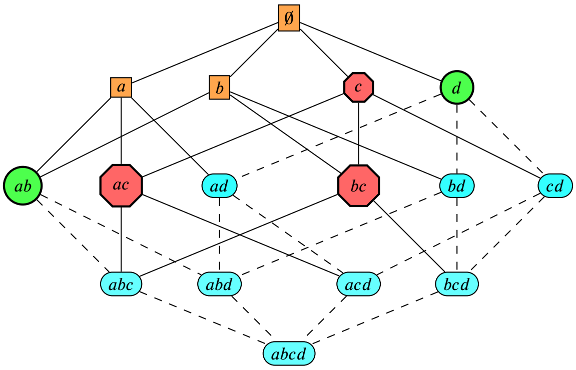

Fig. 10. Lattice space of possible hypotheses for a model with four faults (a, b, c, d) with the computed minimal hypotheses $\Delta_{\min} = \{\{d\}, \{a, b\}\}$. The hypotheses can be partitioned into four subsets: i) the minimal candidates (in circles); ii) the hypotheses that, if they were candidate, would make at least one minimal candidate no longer minimal (in boxes); iii) the hypotheses that, if they were candidate, would be additional minimal candidate (in octagons); iv) the hypotheses that do not change the minimal diagnosis be they candidates or not (in rounded rectangles).

Once again, these results could be extended to non-classical frameworks, but this is outside the scope of this paper.

6.3. Some complexity considerations

With the last improvement now presented, it is possible to give some estimate of the complexity of implementing our framework. In general, finding a single critical observation requires searching over the space of sub-observations to find a minimal sub-observation whose minimal diagnosis equals $\Delta(\mathcal{O})$.

Assuming that this space has a depth of d and a branching factor b, and assuming a greedy search algorithm, this means solving $O(bd)$ diagnosis subproblems (which, in themselves, can be harder to solve than a diagnosis problem for instance for diagnosis approaches that precompute a special structure [10]).

Consider again a diagnosis problem with state-based observations and assume the state-based framework introduced in Subsection 5.1: an observation is a set of variable assignment and the abstraction relation is the subset relation. It is clear that for an observation of size $|b|$, there are 2^b possible sub-observations (subsets) to compute diagnoses for in an exhaustive search. Prior to any optimisation, and noting that it is generally established that the diagnosis procedure of Reiter is exponential in the number of components, a surface analysis would indicate we should expect that exhaustively finding a critical observation will be doubly exponential in the size of the diagnosis problem in the worst case. However, the procedure of Fig. 9 is *not* an exhaustive search, and by making use of the variety of optimisations provided we are able to stay firmly out of the territory of exponential calls to the diagnoser.

Here, we only need to consider b diagnosis sub-problem (Subsection 5.1.2 showed how making use of observable conflicts can reduce this value tremendously). Furthermore, from Subsection 6.2, we know that it is not necessary to solve these sub-problems. Instead, it is possible to determine whether some hypothesis on the faults is consistent with the model and the sub-observations; presumably, an NP-COMPLETE problem.

7. Related work

As far as we are aware, this work is the first one that addresses the problem of finding the part of the observations that supports the diagnosis. Recently, Bertoglio et al. have proposed a notion of “explanatory diagnosis”, in which the explanation is the order in which the faults occurred [21].

Most diagnosis methods, such as proposed by Baroni et al. [12], provide explanations in the form of an assignment of the state variables that is consistent with the observations. In diagnosis of DES, this is a trajectory, a sequence of states and events allowed by the model that produces the observed events. Such explanations illustrate why the diagnoser considers that a given hypothesis is a candidate; however, it does not explain why this is a preferred candidate.

We also see a connection with the work on *alarm filtering*, which consists of handling large volumes of data input and displaying it in a manageable way. Larsson [22] proposed a model-based method to determine which alarms are direct consequences of another alarm. Bauer et al. [23] showed how one can identify which alarms are related to the same incident, such that they can be grouped together.

The most popular question that links diagnosis and observations is that of diagnosability [10,24]. A system is diagnosable if the available sensors are sufficient to determine the occurrence of a fault. If the system is indeed diagnosable, then the diagnosis essentially reduces to a single candidate. Our work takes place in a more general setting in which diagnosability may not hold. One of the main contributions of the present work was to identify that the criterion for deciding if a sub-observation is sufficient is that it should yield the same *minimal* diagnosis.

Similar to our work on reducing the amount of information, other research has proposed the minimisation of the number of sensors whilst still guaranteeing diagnosability [25,26]. The goal here is generally the reduction in the cost of monitoring the system.

Compared to our work, this analysis is performed prior to actual use of the system. Consequently, it needs to be conservative and consider all possible system behaviours, while we only work with a single observation.

Closer to the present work, de Kleer's state-based sequential diagnosis [27], and the event-based variants proposed by Thorsley and Teneketzis, and independently Cassez and Tripakis, introduce the notion of *dynamic observers* which are sensors that can be turned off and on as the system state evolves [28,29]. A dynamic observer is able to determine when a possible observation may become relevant and needs to be monitored. Once again however, this decision has to be made before the observation is produced, which implies that the decision must be conservative.

Also related is older work on the notion of *structural abstraction*, initially proposed by Mozetič, extended by Autio and Reiter, and then Chittaro and Ranon [30–32]. Here a system model starts as very abstracted and is progressively refined with the diagnosis of the previous step informing subsequent steps. This is achieved with simpler abstract component models (super-components, comprising several less abstracted components) which are progressively unpacked until the complex model is reached. Sachenbacher & Struss also tackle a similar problem of composing components and their observations for specific behaviour modelling for domain specific tasks [33]. Crucially, these approaches still explicitly rely on all observations to make a diagnosis, however the idea of reporting the most abstracted model/super-components at which the correct diagnosis is achieved also goes some way towards the same goal of providing decision makers with easier to parse evidence.

Our work requires the ability to diagnose systems with sub-observations, which can be non-trivial, as some diagnosers require a constant monitoring setup. In this scenario it becomes necessary to move to dedicated algorithms [34,35].

7.1. Explanations in artificial intelligence

The topic of explanations in Artificial Intelligence has become very active in the last five to ten years. It is in part driven by the success of deep learning and neural networks which are, for the most part, non-interpretable. There is also a concern that data-driven automated decision makers might reproduce bias and unfairness from existing procedures.

Explanations are sometimes defined as contrastive: why make decision A and not B? This question can be answered by returning a minimal change to the current situation that would yield a better outcome [36].

In the case of a loan denied by a bank for instance, this could lead to the recommendation "Your loan will be accepted if you add a degree to your CV."

In Machine Learning, one popular approach to explanation is to approximate the complex decision function with a function that is similar locally but simpler (for instance, linear). Such an explanation can then be interpreted as follows: "In your context, the important variables are your highest degree and your age" [37,38].

Our work is grounded in symbolic AI and resembles the logical approach developed recently [39,40]. The idea is that a decision or recommendation procedure (such as our diagnoser) is a function f that processes some input (such as our observation) and returns some output (such as our diagnosis). The goal is then to identify what property τ of the input guarantees a certain property p of the output:

$$\forall X \in \mathit{InputSpace}.\ \tau(X) \rightarrow p(f(X)).$$

Our work goes further in several ways. First, our property p is not simply that the output $f(X)$ should be a precise decision or recommendation; it is a property that relates to the diagnosis, and we claim that the relevant property is that the minimal diagnosis $f(X)$ should match the one computed for our observation. Second, unlike the existing approaches, our input space is not limited to vectors of booleans but can be much richer, as in the case of event-based observations. Finally, we identify implementation tricks that helps us converge to τ faster.

We also note that explainable AI planning has become a relevant topic in recent years. Chakraborti *et al.* [41] consider the problem where a human disagrees with the plan computed by an automated planner. They claim that this is as a model reconciliation problem, i.e., that the human and the planner have different models of the world because it is assumed that neither of them cannot make mistakes given a formal planning problem. An explanation is therefore a list of changes that the human should apply to their model. In our setting, explanations are required not because the human and the computer have different models but because the amount of information is too large for the human to process.

8. Conclusion and future works

In this paper, we addressed the problem of finding the part of the observations that is useful for the diagnosis. We defined a *sub-observation* as an abstraction of the observations. We then argued that a sub-observation is *sufficient* if it allows a diagnoser to derive the same minimal diagnosis as the original observations; and we defined *critical observations* as a maximally abstracted sufficient sub-observation. We showed how to compute a critical observation, and discussed a number of algorithmic improvements that also shed light on the theory of critical observations. Finally, we illustrated this framework on both state-based and event-based observations.

We believe that this work is an important landmark for explanatory diagnosis. We envision that this will pave the way to a more interactive relation between the machine (diagnoser) and the human operators in charge of making decisions on the system. We consider for instance that the diagnoser may be incorrect, either because its model is wrong or because it does not have access to as much information as the operator; in both cases, returning a critical observation alongside its diagnosis should aid in the understanding of why the diagnoser was incorrect and how it might be improved.

We also expect that this work can be used to improve privacy. With smart meters and IoT technology, operators now have access to information about users and their behaviours. In an electricity grid for instance, operators are able to tell when users are at home and, sometimes, what electrical devices they are using. Critical observations can be used to obfuscate some of this information, hereby increasing privacy, whilst still maintaining relevant information for monitoring.

We provided two frameworks for state-based and event-based sub-observations, and we hinted at possible extensions of these frameworks. These extensions will need to be formalised, together with the components listed in Subsection 4.3 (sub-observation, sub-observation space, abstraction relation, and the functions sub, MinDiagnose, and $children$). These abstractions include:

- Abstraction of real-valued variables (including time) into qualitative ones.
- Abstraction of part of a message such as identity (for instance, "*Someone* from region R no longer has power") or message content ("User. X was unable to connect to *some* website").
- Partial ordering between observed events.

Another interesting development would be the consideration of an iterative process in which operators can ask for alternate critical observations when they feel unconvinced by the first one. In our theory, we only return the minimal amount of information that entails the minimal diagnosis, but it could include some redundancy.

Finally, an important issue will be to test it on human subjects in realistic settings in order to verify which criteria make this approach the most beneficial in practice. We expect such an evaluation to be a substantial undertaking, with non-trivial models (capturing real systems) and expert involvement.

CRediT authorship contribution statement

Cody James Christopher: Conceptualization, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Alban Grastien:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Both authors have previously worked closely with and collaborated with two editors of AIJ: Sylvie Thiebaux & Patrik Haslum.

Data availability

No data was used for the research described in the article.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artint.2024.104116.

References

[1] R. Reiter, A theory of diagnosis from first principles, Artif. Intell. 32 (1) (1987) 57–95.
[2] J. de Kleer, B. Williams, Diagnosing multiple faults, Artif. Intell. 32 (1) (1987) 97–130.
[3] C.J. Christopher, M.-O. Cordier, A. Grastien, Critical observations in a diagnostic problem, in: 25th International Workshop on Principles of Diagnosis (DX-14), 2014.
[4] C.J. Christopher, M.-O. Cordier, A. Grastien, Critical observations in a diagnostic problem, in: 53rd IEEE Conference on Decision and Control (CDC-14), 2014, pp. 382–387.
[5] C.J. Christopher, A. Grastien, Formulating event-based critical observations in diagnostic problems, in: 26th International Workshop on Principles of Diagnosis (DX-15), 2015, pp. 119–126.
[6] C.J. Christopher, A. Grastien, Formulating event-based critical observations in diagnostic problems, in: 54th IEEE Conference on Decision and Control (CDC-15), 2015, pp. 4462–4467.
[7] C.J. Christopher, Critical observations for model-based diagnosis: Theory and practice, Ph.D. thesis, The Australian National University, November 2019.
[8] G. Lamperti, S. Trerotola, M. Zanella, X. Zhao, Sequence-oriented diagnosis of discrete-event systems, J. Artif. Intell. Res. 78 (2023) 69–141.
[9] J.B. Kruskal, The theory of well-quasi-ordering: a frequently discovered concept, J. Comb. Theory, Ser. A 13 (3) (1972) 297–305.
[10] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, D. Teneketzis, Diagnosability of discrete-event systems, IEEE Trans. Autom. Control 40 (9) (1995) 1555–1575.
[11] Chr Cassandras, St. Lafortune, Introduction to Discrete Event Systems, Kluwer Academic Publishers, 1999.
[12] P. Baroni, G. Lamperti, P. Pogliano, M. Zanella, Diagnosis of large active systems, Artif. Intell. 110 (1) (1999) 135–183.
[13] Ya. Pencolé, M.-O. Cordier, A formal framework for the decentralised diagnosis of large scale discrete event systems and its application to telecommunication networks, Artif. Intell. 164 (1–2) (2005) 121–170.
[14] R. Su, W. Wonham, Global and local consistencies in distributed fault diagnosis for discrete-event systems, IEEE Trans. Autom. Control 50 (12) (2005) 1923–1935.
[15] P. Kan John, A. Grastien, Local consistency and junction tree for diagnosis of discrete-event systems, in: Eighteenth European Conference on Artificial Intelligence (ECAI-08), 2008, pp. 209–213.
[16] G. Jiroveanu, R. Boël, Petri net model-based distributed diagnosis for large interacting systems, in: Sixteenth International Workshop on Principles of Diagnosis (DX-05), 2005, pp. 25–30.

[17] A. Grastien, A. Anbulagan, Diagnosis of discrete event systems using satisfiability algorithms: a theoretical and empirical study, IEEE Trans. Autom. Control 58 (12) (2013) 3070–3083.

[18] P. Haslum, A. Grastien, Diagnosis as planning: two case studies, in: Fifth Scheduling and Planning Applications Workshop (SPARK-11), 2011, pp. 37–44.

[19] A. Grastien, M. Zanella, Discrete-event systems fault diagnosis, in: Fault Diagnosis of Dynamic Systems, Springer, 2019, pp. 197–234.

[20] Al. Grastien, M.-O. Cordier, Chr Largouët, Automata slicing for diagnosing discrete-event systems with partially ordered observations, in: Ninth Congress of the Italian Association for Artificial Intelligence (AI*IA-05), 2005, pp. 270–281.

[21] N. Bertoglio, Gi. Lamperti, M. Zanella, X. Zhao, Explanatory diagnosis of discrete-event systems with temporal information and smart knowledge-compilation, in: Seventeenth International Conference on the Principles of Knowledge Representation and Reasoning (KR-20), 2020, pp. 130–140.

[22] J.E. Larsson, Real-time root cause analysis with multilevel flow models, in: 20th International Workshop on Principles of Diagnosis (DX-09), 2009, pp. 3–8.

[23] A. Bauer, A. Botea, A. Grastien, P. Haslum, J. Rintanen, Alarm processing with model-based diagnosis of discrete event systems, in: AI for an Intelligent Planet (AIIP-11), 2011, pp. 1–8.

[24] M.-O. Cordier, P. Dague, F. Lévy, J. Montmain, M. Staroswiecki, L. Travé-Massuyès, Conflicts versus analytical redundancy relations: a comparative analysis of the model based diagnosis approach from the artificial intelligence and automatic control perspectives, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 34 (5) (2004) 2163–2177.

[25] L. Brandán Briones, A. Lazovik, Ph. Dague, Optimal observability for diagnosability, in: Nineteenth International Workshop on Principles of Diagnosis (DX-08), 2008, pp. 31–38.

[26] J. Armengol, A. Bregón, T. Escobet, E. Gelso, M. Krysander, M. Nyberg, X. Olive, B. Pulido, L. Travé-Massuyès, Minimal structurally overdetermined sets for residual generation: a comparison of alternative approaches, in: Seventh IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes (SafeProcess-09), 2009, pp. 1480–1485.

[27] J. de Kleer, Using crude probability estimates to guide diagnosis, Artif. Intell. 45 (3) (1990) 381–391.

[28] D. Thorsley, D. Teneketzis, Active acquisition of information for diagnosis and supervisory control of discrete event systems, J. Discrete Event Dyn. Syst. 17 (2007) 531–583.

[29] Fr. Cassez, St. Tripakis, Fault diagnosis with static or dynamic diagnosers, Fundam. Inform. 88 (4) (2008) 497–540.

[30] I. Mozetič, Hierarchical model-based diagnosis, Int. J. Man-Mach. Stud. 35 (3) (1991) 329–362.

[31] K. Autio, R. Reiter, Structural abstraction in model-based diagnosis, in: ECAI '98, 1998, pp. 269–273.

[32] L. Chittaro, R. Ranon, Hierarchical model-based diagnosis based on structural abstraction, Artif. Intell. 155 (1–2) (2004) 147–182.

[33] M. Sachenbacher, P. Struss, Task-dependent qualitative domain abstraction, Artif. Intell. 162 (1–2) (2005) 121–143.

[34] Al. Grastien, Diagnosis of hybrid systems with SMT: opportunities and challenges, in: 21st European Conference on Artificial Intelligence (ECAI-14), 2014, pp. 405–410.

[35] X. Su, M. Zanella, Al. Grastien, Diagnosability of discrete-event systems with uncertain observations, in: 25th International Joint Conference on Artificial Intelligence (IJCAI-16), 2016, pp. 1265–1271.

[36] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[37] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-16), 2016, pp. 1135–1144.

[38] Sc. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: 31st Advances in Neural Information Processing Systems (NeurIPS-17), 2017, pp. 4765–4774.

[39] A. Ignatiev, N. Narodytska, J. Marques-Silva, On relating explanations and adversarial examples, in: 33rd Advances in Neural Information Processing Systems (NeurIPS-19), 2019, pp. 15857–15867.

[40] Ad. Darwiche, Ch. Ji, On the computation of necessary and sufficient explanations, in: 36th Conference on Artificial Intelligence (AAAI-22), vol. 36, 2022, pp. 5582–5591.

[41] T. Chakraborti, S. Sreedharan, Yu. Zhang, S. Kambhampati, Plan explanations as model reconciliation: moving beyond explanation as soliloquy, in: 26th International Joint Conference on Artificial Intelligence (IJCAI-17), 2017, pp. 156–163.