Explications de diagnostic à base de modèle

Alban Grastien
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

alban.grastien@cea.fr

Résumé

Diagnostiquer consiste à détecter et identifier des défauts ou pannes dans un système complexe — réseau électrique ou de communication, chaîne de production, etc. Le diagnostiqueur est parfois incorrect en raison d'approximations inéluctables dans le modèle du système. Dans ce travail, nous proposons d'adjoindre au diagnostic une justification du résultat sous la forme d'une synthèse des observations reçues. Ainsi un opérateur peut vérifier la validité du diagnostic. Cette justification doit rester simple pour être compréhensible et nous faisons le lien avec les « chroniques », développées dans les années 90 pour le diagnostic.

Mots-clés

Diagnostic à base de modèle. IA explicative.

1 Introduction

Le diagnostic est la branche de l'IA qui s'intéresse à détecter, isoler et identifier l'occurrence de fautes dans les systèmes partiellement observables. Nous nous intéressons ici au diagnostic à base de modèle dont le raisonnement se base sur une comparaison entre les observations espérées et celles réellement obtenues.

Nous nous plaçons principalement dans un contexte industriel (chaîne de production, réseau électrique ou de communication), même si les techniques de diagnostic peuvent s'appliquer à une large panoplie de systèmes. Un frein important au développement du diagnostic en industrie est le caractère « boîte noire » de l'approche et le risque d'erreurs. Ce risque est particulièrement important car le modèle nécessite souvent des approximations voire inclut des erreurs dû à la difficulté de collecter des informations pour les systèmes complexes.

Étant donné ce constat, nous voyons le diagnostiqueur comme un outil d'aide à la décision plutôt qu'un décideur final que l'utilisateur humain doit suivre aveuglément. Pour ce faire, nous souhaitons étendre le diagnostiqueur pour le faire renvoyer, en plus du diagnostic, une *explication* qui aide l'utilisateur à suivre le raisonnement suivi par celui-ci. Ainsi, en plus d'aider à la compréhension finale du diagnostic et de ses répercutions, l'explication doit permettre à l'utilisateur de déterminer si le diagnostiqueur a commis une erreur. Ce second objectif est très important pour nous : nous pensons que le diagnostiqueur a de grandes chances de commettre des erreurs, et que ces erreurs pourraient être désastreuses si elles conduisent à une mauvaise réponse de la part de l'utilisateur. Il est donc important de s'assurer que ces explications sont « simples ». En ce sens, un diagnostic qui ne serait pas accompagné d'une explication simple serait inutile.

Dans cet article, nous définissons une explication de diagnostic comme un objet mathématique (par exemple, un prédicat) qui permet à l'utilisateur de produire le diagnostic. Si la validité du prédicat et l'inférence du diagnostic à partir de ce prédicat sont toutes deux « simples », alors l'explication est compréhensible pour un utilisateur humain. S'il existe une explication simple à tous les comportements fautifs, on dit que le système est explicable.

Dans un second temps, nous proposons une implémentation d'explication sous forme de « chronique ». Une chronique est un motif d'événements observables. L'usage des chroniques a été proposé dans les années 90 [7] pour le diagnostic : une ou plusieurs chroniques sont construites pour chaque faute et la tâche de diagnostic se résume à identifier les chroniques dans le flot d'observation. L'avantage des chroniques est qu'elles peuvent être utilisées sans modèle car les chroniques peuvent être construites à partir de connaissance experte. D'autre part, des algorithmes efficaces ont été développés [6].

Nous représentons donc les explications comme des chroniques. Ceci permet de réhabiliter les chroniques en montrant qu'une qualité de celles-ci, qui n'était pas mentionnée dans la litérature scientifique, est qu'elles sont interprétables par les utilisateurs.

Un autre problème des systèmes de diagnostic à base de chronique est qu'il est parfois nécessaire de construire de nombreuses chroniques qui plus est de grande taille. Il est parfois même impossible de construire un ensemble de chroniques complet quand bien même chaque faute est diagnosticable. Nous montrons cependant que si le système est explicable — ce que nous considérons comme une condition indispensable pour un système à diagnostiquer — alors le nombre de chroniques est nécessairement faible et celle-ci sont petites (puisque chaque comportement a une explication simple). Nous en concluons qu'une approche à base de chronique est viable.

Le reste de cet article est divisé comme suit. La prochaine section présente les définitions classiques de diagnostic. La notion d'explication est décrite dans la section 3 et la question d'explicabilité dans la section 4. Finalement, nous discutons de travaux similaires avant de conclure.

2 Contexte

2.1 Diagnostic de SÉD

Un SÉD (système à événements discrets) est un modèle de système dynamique. Nous reprenons une définition proche de celle proposée initialement par Sampath et co-auteurs [15]. Un *SÉD* est un tuple $M = \langle Q, \Sigma, T, I, \Sigma_o, \Sigma_F \rangle$ tel que Q est un ensemble d'*états* dont $I \subseteq Q$ le sous-ensemble des *états initiaux*, Σ est l'ensemble des *événements* dont Σ_o le sous-ensemble d'*événements observables* et $\Sigma_F \subseteq \Sigma$ le sous-ensemble d'*événements de faute* et $T \subseteq Q \times \Sigma \times Q$ est l'ensemble des *transitions*. Pour simplifier la discussion, nous considérons qu'il n'y a qu'un seul événement de faute ($\Sigma_F = \{e_F\}$), mais ces travaux se généralisent facilement à un non singleton.

La sémantique du SÉD repose sur la notion de chemin. Un *chemin* $\rho \in Chs(M)$ de longueur k est une double séquence $q_0 \xrightarrow{e_1} q_1 \xrightarrow{e_2} \dots \xrightarrow{e_k} q_k$ d'états et de événements telle que $q_0 \in I$ est un état initial et pour chaque $i \in \{1, \dots, k\}$, $\langle q_{i-1}, e_i, q_i \rangle \in T$ est une transition. La restriction de la séquence d'événements $[e_1, \dots, e_k]$ aux événements observables est appelé une *trace*. Un chemin comprenant un événement de faute ($\exists i \in \{1, \dots, k\}.\ e_i = e_F$, noté $e_F \in \rho$) est qualifié de *fautif* ; sinon, le chemin est *normal*. S'il y a au moins d événements après la faute ($i \leq k - d$), on parle de chemin d-fautif.

Un *problème de diagnostic* est une paire $P = \langle M, O \rangle$. Le diagnostic est le problème consistant à déterminer si la faute e_F a eu lieu. Formellement, le *diagnostic* de O est défini par

$$\Delta_M(O) = \begin{cases} \text{N} & \text{si } \exists \rho \in Chs(M). \\ & \quad obs(\rho) = O \wedge e_F \notin \rho, \\ \text{F} & \text{sinon.} \end{cases}$$

Les symboles N et F signifient respectivement « normal » et « fautif ».

On notera que la définition de diagnostic est « optimiste » en ce sens que le diagnostiqueur se contente de trouver une trace nominale en accord avec les observations pour renvoyer un diagnostic normal. Cette définition est acceptable dans la mesure où nous faisons une seconde hypothèse, à savoir que le système est diagnosticable, c'est-à-dire que nous supposons qu'il existe un délai borné à l'issue duquel toute faute sera détectée. Formellement, la *diagnosticabilité* pour un délai $d \in \mathbf{N}$ est la propriété

$$\forall \rho \in Chs(M).\ \rho \text{ est } d\text{-fautif} \Rightarrow \Delta_M(obs(\rho)) = \text{F}.$$

Exemple *Considérons l'exemple de la figure 1 (gauche). Le chemin* $\rho = 0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{b} 1' \xrightarrow{c} 1 \xrightarrow{a} 2 \xrightarrow{e_F} 3 \xrightarrow{a} 4$ *produit la trace* $O = obs(\rho) = aabcaa$. *Le diagnostic de cette trace est* $\Delta_M(O) = \text{F}$.

2.2 Chroniques

Les chroniques sont un formalisme de système expert développé dans les années 90 pour diagnostiquer les SÉD de manière efficace et sans qu'il soit nécessaire de construire un modèle. Une chronique décrit un motif d'événements symptomatique d'une faute.

Un *graphe orienté* est une paire $G = \langle V, E \rangle$ telle que V est un ensemble de *nœuds* et $E \subseteq V \times V$ un ensemble d'*arêtes*. Un *cycle* dans G est une séquence de $k > 1$ nœuds $v_1, \dots, v_k$ qui satisfait les deux propriétés $\forall i \in \{1, \dots, k-1\}.\ \langle v_i, v_{i+1} \rangle \in E$ et $v_1 = v_k$. Un *graphe orienté acyclique* (GOA) est un graphe orienté qui ne comporte aucun cycle ; en particulier, il n'admet aucune boucle ($\forall v \in V.\ \langle v, v \rangle \notin E$).

Étant donné l'ensemble d'événements observables Σ_o, une chronique (atemporelle) [7] est un triplet $\theta = \langle G, L_V, L_E \rangle$ tel que $G = \langle V, E \rangle$ est un GOA, $L_V : V \to \Sigma_o$ et $L_E : E \to 2^{\Sigma_o}$ sont deux fonctions qui associent, respectivement, l'ensemble des nœuds et l'ensemble des arêtes à, respectivement, un événement observable et un ensemble d'événements observables. [1]

On appelle les événements mentionnés par L_V des *observations nécessaires* et les événements de L_E des *observations interdites*. Une chronique est reconnue si les observations nécessaires apparaissent dans la trace tandis que les événements interdits n'y apparaissent pas, ceci dans l'ordre précisé par le graphe. Les événements interdits représentent le fait que l'absence de certains événements est inattendue ; ainsi, si un événement et son annulation (par exemple, ouverture/fermeture ou entrée/sortie) sont observables, on s'attendra à voir une annulation entre deux occurrences du premier événement : ouverture, suivi de fermeture, suivi d'ouverte. C'est plus ou moins ce que les événements a et b représentent dans l'exemple de la figure 1 (gauche).

Formellement, étant données une trace comportant n événements observables $O = o_1, \dots, o_n$ et une chronique $\theta = \langle G, L_V, L_E \rangle$, une *(fonction de) correspondance* entre O et θ (indiquant quand, dans la trace O, les événements nécessaires de θ ont lieu) est une fonction $t : V \to \{1, \dots, n\}$ qui satisfait les propriétés suivantes :

1. $\forall v_1, v_2 \in V.\ v_1 \neq v_2 \Rightarrow t(v_1) \neq t(v_2)$,
2. $\forall \langle v_1, v_2 \rangle \in E.\ t(v_1) < t(v_2)$,
3. $\forall v \in E.\ o_{t(v)} = L_V(v)$ et
4. $\forall \langle v_1, v_2 \rangle \in E.\ \forall i \in \{t(v_1) + 1, \dots, t(v_2) - 1\}.$ $o_i \notin L_E(\langle v_1, v_2 \rangle)$.

S'il existe une correspondance entre O et θ, on dit que θ est une chronique de O et que O suit θ. On utilise aussi la notation $O \in [[\theta]]$.

Lien avec le diagnostic La chronique θ est *typique de la faute* e_F (ou, simplement, *typique*) si toute trace O cohérente avec le modèle M et qui suit θ provient d'une chemin fautif :

$$\forall \rho \in Chs(M).\ obs(\rho) \in [[\theta]] \Rightarrow e_F \in \rho.$$

En conséquence de quoi, toute trace qui suit θ peut être diagnostiquée comme fautive :

Lemme 1 *Si la chronique* θ *est typique, alors le résultat suivant est correct*

$$O \in [[\theta]] \Rightarrow \Delta_M(O) = \text{F}.$$

1. Les définitions classiques ajoutent des informations temporelles aux arêtes du graphe ; étendre les résultats pour ce type de définition est un des axes de recherche futurs de nos travaux.

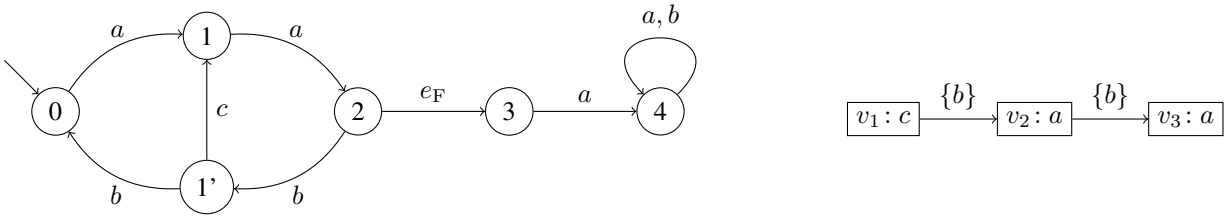

FIGURE 1 – Gauche : Exemple de SÉD diagnosticable si l'ensemble Σ_o comporte a et b. Il est explicable (cf. Section 4) pour une taille d'explication de $S = 3$ si Σ_o vaut $\{a, b, c\}$; sinon, il n'est pas explicable. Droite : exemple de chronique typique si a, b et c sont observables.

Exemple *Reprenons l'exemple de la figure 1 et considérons la chronique sur la droite de la figure. Cette chronique comprend trois nœuds associés aux événements nécessaires c, a et a respectivement. L'arête entre v_2 et v_3 comporte également un événement interdit b. On peut facilement démontrer que cette chronique est typique puisqu'elle trahit la séquence de transitions $1' \xrightarrow{c} 1 \xrightarrow{a} 2 \xrightarrow{e_F} 3 \xrightarrow{a} 4$. Ainsi, une trace qui suit cette chronique peut facilement être diagnostiquée quelque soit sa longueur.*

Un système de reconnaissance de chroniques (SRC ou *chronicle recognition system* CRS en anglais) est un diagnostiqueur qui utilise un ensemble Θ de chroniques typiques. Le SRC lit la trace fournie par le système diagnostiqué et vérifie si celle-ci suit une des chroniques. Tant que ceci n'arrive pas, le SRC renvoie N ; mais dès qu'une des chroniques est suivie par la trace, alors le diagnostiqueur renvoie F. On notera qu'un tel diagnostiqueur n'est pas toujours exact mais ses approximations peuvent être acceptables sous certaines conditions. Par exemple, un tel diagnostiqueur introduit parfois des délais entre la détection d'une faute et la date à laquelle on disposait de suffisamment d'informations pour la détection. Si ce délai est borné et si cette borne est suffisamment faible, alors ce délai n'est pas problématique en pratique (on se rappellera qu'il y a déjà un délai entre l'occurrence de la faute et la date à laquelle celle-ci peut-être détectée).

Un diagnostiqueur à base de chronique est surtout très facile à implémenter et très efficace. Il est donc intéressant de pouvoir remplacer un diagnostiqueur complet tel que [15, 13, 19] pour un simple SRC. S'il existe un ensemble fini et « petit » Θ de chroniques tel que toute occurrence de faute sera détectée par un SRC muni de Θ, on dit que Θ couvre les comportements de faute du système. On dit alors que le diagnostic à base de chronique est une méthode *viable* pour le SÉD M.

3 Explications

Nous sommes intéressés par le calcul d'explications pour le diagnostic. De manière informelle, une explication est une information permettant à un utilisateur humain de comprendre le raisonnement du diagnostiqueur et de se convaincre de la correctitude du diagnostic. Le cas échéant, une explication pourrait être utilisée pour détecter une erreur de la part du diagnostiqueur. En effet, même si le diagnostiqueur est correctement implémenté, la théorie du diagnostic repose sur des hypothèses dont on peut légitimement contester la validité. Ainsi, le diagnostic à base de modèle présume l'existence d'un modèle (le SÉD) correct. Or, pour de nombreuses raisons, cette présomption peut être malvenue :

— Le modèle peut contenir des erreurs. Par exemple, nous avons pu constater lors de nos collaborations avec certaines compagnies de distribution d'électricité que celles-ci ne disposaient pas de base de données centralisées de leur réseau électrique, ce qui rendait toute modélisation du réseau approximative *a minima*.

— Le modèle peut ignorer certaines situations exceptionelles requérant une modélisation plus fine. Pour reprendre l'exemple du réseau de distribution d'électricité, un événement sportif ou culturel inhabituel pourra modifier les paramètres habituels du réseau et tromper le diagnostiqueur. Une explication du diagnostic devrait permettre à l'opérateur en charge du réseau de comprendre que la panne prédite est possiblement liée à ces conditions inhabituelles et l'encourager à chercher d'autres symptômes avant d'agir prématurément.

3.1 Qu'est-ce qu'une explication

Nous considérons donc qu'une explication est une information fournie en parallèle du diagnostic qui permet — d'après l'agent en charge de générer l'explication — d'arriver à la même conclusion que le diagnostiqueur mais avec un effort de raisonnement plus faible. Par exemple, une explication peut pointer les observations importantes dans la trace fournie par le système.

Nous proposons de définir une explication comme suit :

Définition 1 *Étant donné un problème de diagnostic $P = \langle M, O\rangle$ où M est un SÉD et O est une trace telle que $\Delta_M(O) = \text{F}$, une* explication *est un object mathématique x tel que les assertions suivantes sont vraies :*

1. *x est une propriété de O, ce qu'on notera $O \models x$;*
2. *x est une preuve du diagnostic, à savoir que*

$$\forall \rho \in \mathit{Chs}(M).\ (\mathit{obs}(\rho) \models x) \Rightarrow \Delta_M(\mathit{obs}(\rho)) = \text{F},$$

ce que l'on notera $x \models \text{F}$.

Une explication est simple *s'il est simple de prouver $O \models x$ et $x \models \text{F}$.*

En pratique, on n'est intéressé que par des explications simples.

Il convient de faire plusieurs remarques sur cette définition. Tout d'abord, elle ne donne pas de définition exacte d'explication. C'est intentionel parce que le but est justement de fournir une définition générale qui permette à différents chercheurs de proposer différentes définitions. Nous donnons ci-après une définition basée sur les chroniques mais d'autres définitions seraient possibles. Par ailleurs, la définition est ici donnée pour des problèmes de diagnostic de SÉD mais elle s'applique à n'importe quel type de formalisme. Ensuite, la définition reste informelle sur ce qu'on entend par « simple ». C'est notamment lié au fait que la définition d'explication reste ouverte. C'est aussi parce que les critères précis et spécifiques déterminant ce qui rend une explication « simple » devront être définis au cas par cas. Est-ce la taille de l'explication qui compte ? La taille du modèle qu'il faut considérer ? Le type de raisonnement à appliquer ? etc. Enfin, on notera que la définition présume que le diagnostic est fautif ($\Delta_M(O) = \text{F}$). La raison est qu'un comportement nominal est généralement plus difficile à prouver qu'un comportement fautif parce qu'il implique l'absence d'observations anormales ; la raison pour laquelle un comportement est nominal est que *toutes les observations* restent dans le périmètre décrit par le modèle. [2]

On peut donc voir une explication comme un prédicat satisfait par la trace tel que, d'une part, la vérification que la trace satisfait ce prédicat est facile et, d'autre part, l'occurrence de faute se déduit facilement depuis le prédicat. Ainsi par exemple, un prédicat qui capturerait l'information « J'ai tourné la clef mais le moteur n'a pas démarré » indique, *a priori*, une panne. Dans cet exemple, l'explication ressemble à une chronique, ce qui n'est pas anodin : nous proposons en effet d'utiliser des chroniques pour représenter des explications. La définition est cependant ouverte. Ainsi, si le SÉD est représenté de manière symbolique (à savoir que les états et transitions ne sont pas listées explicitement mais de manière implicite via des variables d'états), l'explication peut inclure des faits (prédicats sur les variables d'états) qui peuvent être dérivés à partir des observations.

3.2 Explications et chroniques

Nous proposons d'instantier la définition 1 avec une chronique typique. L'idée est que la difficulté du diagnostic pour un humain est le plus souvent le volume des observations (taille de la trace). Le rôle de l'explication est d'extraire les observations (ou leur absence) importantes de la trace pour les mettre en avant. Vérifier qu'une trace suit une chronique est facile (d'autant plus que nous demandons que la correspondance soit incluse dans l'explication), et nous partons de l'hypothèse que prouver l'occurrence de la faute à partir d'une chronique typique (si celle-ci est suffisamment petite) est également facile, au moins pour un utilisateur expert.

2. Pour un système diagnosticable, on ne peut généralement prouver la nominalité du comportement que jusqu'à la date actuelle moins d, le délai de diagnosticabilité ; on pourrait donc également définir une explication pour la nominalité du comportement jusqu'à cette date.

Définition 2 *Étant donné un problème de diagnostic $P = \langle M, O\rangle$ où M est un SÉD et O est une trace telle que $\Delta_M(O) = \text{F}$, une* explication à base de chronique *est une paire $x = \langle \theta, t\rangle$ telle que θ est une chronique typique de O et t est une correspondance entre O et θ.*

L'explication comprend donc une chronique typique qui est la condition suffisante pour diagnostiquer la faute ($x \models \text{F}$). De plus, l'explication contient une correspondance entre O et θ qui prouve — et facilite la preuve — que O suit θ ($O \models x$). On dira parfois que θ explique la faute, en ce sens qu'il existe une correspondance t telle que $\langle \theta, t\rangle$ est une explication.

Définition 3 *La* taille *d'une chronique θ, notée $|\theta|$, est le nombre de nœuds dans le graphe. Un* seuil de complexité *est un entier $S \in \mathbf{N}$. Étant donné un seuil de complexité $S \in \mathbf{N}$, la chronique est* simple *si sa taille est au plus S : $|\theta| \leq S$.*

Dans la définition 3, on utilise la taille de la chronique, c'est-à-dire son nombre de nœuds, comme proxy pour la complexité. On pourrait choisir d'autres mesures telles que le nombre de composants qu'elle mentionne. Cette définition reste donc ouverte.

On fait l'hypothèse suivante :

Hypothèse 1 *Une explication à base de chronique $x = \langle \theta, t\rangle$ est simple si θ est simple.*

On note en particulier que, étant donnée la correspondance t entre O et θ, vérifier $O \models x$ est trivial puisqu'il suffit de vérifier que les observations nécessaires sont présentes aux dates indiquées tandis que les observations interdites sont, quant à elles, absentes. La seconde condition, que la preuve $\theta \models \text{F}$ est simple, est loin d'être évidente ; c'est pour cela que nous parlons ici d'« hypothèse ». Cependant, nous considérons qu'une chronique simple devrait naturellement capturer les informations *évidemment* suffisantes pour conclure à une faute. Par ailleurs, en pratique, on pourra imaginer une procédure incrémentale et interactive dans laquelle un utilisateur pourra demander une explication différente ou plus complète. Ceci est une piste de recherche future.

Les explications à base de chronique peuvent être vues comme un cas spécial d'observations critiques [5]. Ces dernières sont définies comme une abstraction de la trace qui permettent de produire le même (minimal) diagnostic, c'est-à-dire que toutes les informations redondantes ou non pertinentes de la trace ont été supprimées. Nous avons choisi ici de privilégier le lien avec les chroniques à cause de l'existence des SRC qui offrent l'explication de manière immédiate sans nécessiter de traitement post-diagnostic.

Exemple *Appliquons les définitions précédentes au problème de diagnostic appliqué au SÉD de la figure 1 (gauche) pour la trace $O = aabcaabab$. Le diagnostic est $\Delta_M(O) = \text{F}$. Une explication pour ce diagnostic est $\langle \theta, f\rangle$ où θ est la chronique de la figure 1 (droite) et $t = \{c \mapsto 4, a \mapsto 5, a \mapsto 6\}$ est la correspondance entre θ et O.*

4 Explicabilité

Nous cherchons maintenant à déterminer si un SÉD est explicable, c'est-à-dire s'il est construit de telle manière que non seulement toute faute sera diagnostiquée mais également que ce diagnostic pourra être expliqué. Si le système n'est pas explicable, on pourra alors ajouter des capteurs qui donneront plus d'options pour générer une explication simple. On pourra aussi modifier le comportement du système (par exemple via ses contrôleurs) de telle manière que celui-ci renvoie des indices permettant d'expliquer le diagnostic.

4.1 SÉD non explicables

Puisque la simplicité d'une explication repose sur sa taille, on définit la notion de complexité d'explicabilité d'une trace comme étant la taille de sa plus petite explication. On ne cherchera pas forcément à trouver cette explication (ou une de ces explications), mais cette notion est utile pour décider l'existence d'une explication simple.

Définition 4 *Étant donnés un SÉD M et une trace O telle que $\Delta_M(O)$ est* F*, la* complexité d'explicabilité *de O est la taille de sa plus petite explication :*

$$comp(O) = \min_{O \models \langle \theta, t \rangle \models \mathrm{F}} |\theta|.$$

Nous illustrons la question d'explicabilité sur un exemple avant de fournir une définition formelle.

Exemple *Nous reprenons l'exemple de la figure 1 mais nous considérons que c n'est plus observable ($\Sigma_o = \{a, b\}$). La trace $O = aabcaaba$ que nous considérions jusqu'alors est maintenant $O' = aabaaba$. Le diagnostic de cette trace est toujours le même : $\Delta_{M'}(O') =$* F*. La plus simple chronique qu'on puisse construire et qui sert d'explication pour O' est illustrée sur la figure 2. Si le seuil de complexité S est inférieur à 5, alors cette chronique n'est pas simple et il n'y a pas de simple explication pour la trace O'.*

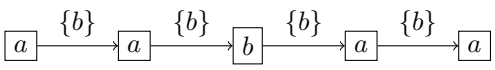

FIGURE 2 – Une chronique non simple pour le système de la figure 1.

Cet exemple nous conduit à la définition suivante :

Définition 5 *Étant donné un seuil de complexité S et un SÉD M diagnosticable avec un délai d, M est (d, S)-explicable si pour tout chemin d-fautif ρ, la complexité d'explicabilité de sa trace est S ou moins.*

Exemple *Selon la définition 5, l'exemple de la figure 1 n'est pas explicable lorsque c n'est pas observable. En réalité, toute trace fautive de M peut être expliquée (et même, étant donné notre définition d'explication, il est toujours possible de créer une explication pour n'importe quel système diagnosticable parce qu'on peut créer une chronique qui corresponde précisément à la trace), mais ces explications peuvent être de longueur rédhibitoire.*

En revanche, dans le même exemple pour $\Sigma_o = \{a, b, c\}$, le système est explicable. En effet, deux chroniques sont suffisantes pour couvrir tous les cas de fautes : la première est celle de la figure 1 (droite), la seconde est identique si ce n'est que l'événement observable c du premier nœud est remplacé par l'événement a.

Il est possible de prouver que la notion d'explicabilité satisfait des propriétés de monotonicité telles que celle-ci :

Lemme 2 *Soient $M = \langle Q, \Sigma, T, I, \Sigma_o, \Sigma_F \rangle$ et $M' = \langle Q, \Sigma, T, I, \Sigma'_o, \Sigma_F \rangle$ deux SÉD tels que $\Sigma_o \subseteq \Sigma'_o$. Si M est (d, S)-explicable, alors pour tout $d' \geq d$, M' est (d', S)-explicable.*

Le lemme 2 peut se démontrer facilement par le fait que toutes les explications simples trouvées pour M sont également des explications simples pour M'. Augmenter la taille de d ou le nombre d'événements observables ne peut qu'augmenter l'espace des explications.

Notons cependant que le lemma 2 repose sur un certain nombre d'hypothèses implicites. Tout d'abord, il considère que le système est d-diagnosticable. En effet, si le système n'était pas diagnosticable, alors on pourrait faire face à des scénarios bizarres. Ainsi, il serait possible que tous les comportements de M qui sont diagnosticables soient explicables, tandis que certains comportements de M' (qui sont diagnosticables pour M' mais pas pour M) seraient non explicables. Autrement dit, le diagnostiqueur deviendrait plus puissant (capable de diagnostiquer plus de comportements fautifs) mais ne serait pas capable d'expliquer le diagnostic de ces comportements.

Par ailleurs, on notera l'importance du fait que la simplicité de l'explication est basée uniquement sur la taille de la chronique et non pas sur la fonction de correspondance. Et en effet, imaginons une autre définition de la taille de l'explication. Dans cette définition, le but est de pouvoir afficher sur un écran tous les événements observés pendant la chronique. On peut afficher autant d'événements observés que de lignes. La taille d'une explication devient la différence entre l'index du dernier événement mentionné par la fonction de correspondance et celui du premier. On voit alors qu'ajouter des événements à l'ensemble Σ_o *pollue* l'écran et rend l'explication non simple.

Il convient donc de faire attention à la définition d'explication et de simplicité.

Pour finir, nous montrons qu'il existe des exemples pour lesquels il n'y a aucun seuil de complexité permettant de garantir l'explicabilité.

Définition 6 *Pour un SÉD d-diagnosticable, une famille de chemins* divergente *est une séquence infinie de chemin $\rho_1, \rho_2, \ldots$ telle que chaque chemin ρ_i est d-fautif et la complexité d'explicabilité augmente de manière strictement monotone :*

$$\forall i \in \mathbf{N}.\ comp(obs(\rho_i)) < comp(obs(\rho_{i+1})).$$

Exemple *Nous revenons à l'exemple de la figure 1 où $\Sigma_o = \{a, b\}$ et considérons la trace fautive $\rho_n = a(ab)^n aa$. La plus petite chronique pour cette trace est similaire à la figure 2 et comporte $3 + 2n$ nœuds, ce qui signifie que $\rho_1, \rho_2, \ldots$ forme une famille de chemins divergente. On conclue donc que, bien que diagnosticable, il n'existe aucun seuil qui rende ce système explicable.*
Puisque la complexité est un entier, une famille de chemins divergente dans un SÉD démontre qu'il n'y a pas de borne maximale pour la plus petite explication d'une trace de ce SÉD.

Lemme 3 *Soient m, ℓ, et d trois entiers. Il n'existe pas de fonction ub tel que pour tout SÉD d-diagnosticable comportant m états et ℓ événements et tout chemin ρ d-fautif, il existe une explication de taille au plus $ub(m, \ell, d)$.*

Ce lemme est prouvé par l'exemple précédent. En revanche, on peut prouver le lemme suivant.

Lemme 4 *Étant donné un SÉD d-diagnosticable comportant m états, si il existe un chemin d-fautif dont l'explication la plus courte a une taille supérieure à 2^m, alors le SÉD a une famille de chemins divergente.*

La preuve du lemme 4 repose sur l'idée suivante. Considérons un chemin ρ et sa trace $obs(\rho)$, et considérons que sa plus courte explication utilise la chronique θ de taille supérieure à 2^m. On peut d'ailleurs supposer que les arêtes de la chronique définissent un ordre total sur les nœuds du graphe parce que les chroniques basées sur des ordres strictement partiels peuvent être interprétées comme des unions de chroniques totalement ordonnées. De manière générale, il est possible d'associer à chaque nœud v_j d'une chronique un ensemble de croyance $\mathcal{Q}$ qui représente l'ensemble des états dans lequel le système pourrait se trouver étant donné le préfixe de ce nœud. Si le même ensemble de croyance $\mathcal{Q}$ est associé à de nœuds v_j, v_k, alors le raisonnement a effectué une boucle. Il est possible de modifier ρ pour dupliquer cette boucle de telle manière que la chronique θ n'est plus appropriée mais qu'il faille la remplacer par une autre chronique θ' plus longue (dans laquelle la boucle a été dupliquée le même nombre de fois). Il est possible qu'une autre chronique de longueur identique existe qui explique à la fois ρ et ρ', mais si on applique le même raisonnement et si on duplique la boucle dans ρ' pour chacune des chroniques, on obtient invariablement après un nombre fini de modifications un nouveau chemin ρ^k dont la complexité d'explicabilité est supérieure à celle de ρ. En appliquant cette approche récursivement, on génère une famille de chemins divergente.
Ce lemme démontre donc qu'on peut vérifier si le SÉD contient une famille de chemins divergente en vérifiant si l'ensemble des chroniques typiques de taille 2^m couvre tous les chemins d-fautifs, même si la taille des chroniques rend impossible cette vérification en pratique.

4.2 Explicabilité et viabilité du SRC

On peut à présent définir le problème d'explicabilité.

Définition 7 *Étant donnés une valeur d, un SÉD M qui est d-diagnosticable et un seuil de complexité S, le problème d'explicabilité consiste à déterminer si M est (d, S)-explicable.*

Notre définition d'explicabilité repose sur les chroniques. À supposer que l'on considère que le seuil S est bas, alors le nombre de chroniques qu'on peut construire est aussi faible. Essentiellement, il est de l'ordre de $\ell^S \times (2^\ell)^S$ où le premier terme indique le nombre possible de combinaisons d'événements nécessaires tandis que le second terme fait référence aux événements interdits. En pratique cependant, nous pensons que le nombre de chroniques dont un SRC aura besoin pour diagnostiquer correctement le système sera beaucoup plus faible. (Ou alors, on pourra ajuster la définition de chronique. Par exemple, l'assertion « l'événement observable e est suivi de k événements observables différents de e' » ne s'exprime pas facilement avec les chroniques parce que un nœud est associé avec un seul événement ; à la place, on pourrait associer les nœuds avec un ensemble d'événements pour éviter ce cas de figure.)
Ceci nous autorise à produire le résultat suivant :

Lemme 5 *En première approximation, si un SÉD est explicable pour un seuil de complexité faible, alors le diagnostic à base de chronique est une approche viable pour le diagnostic de ce SÉD.*

Nous considérons que ce résultat est très intéressant parce qu'il révèle une propriété intéressante du diagnostic par chronique : à savoir que non seulement il produit de manière naturelle des diagnostics explicables, mais que pour tout problème explicable, l'approche par chronique est applicable. Ainsi, si l'explicabilité est une contrainte utilisateur forte, il est possible d'utiliser un SRC pour implémenter le diagnostic. Étant donné la simplicité des SRC, c'est un résultat très encourageant.

4.3 Vérifier l'explicabilité

Nous souhaitons maintenant vérifier si un système est explicable et, par la même occasion, générer un ensemble de chroniques typiques qui couvrent tous les cas fautifs prédits par le modèle.
Tout d'abord, déterminer si une chronique est typique est relativement simple à implémenter. En effet, une chronique représente un langage, le langage de toutes les séquences d'événements observables qui suivent cette chronique. Ce langage est régulier et facile à représenter. Vérifier si la chronique est typique consiste à effectuer un produit synchrone de ce langage avec le modèle pour vérifier s'il existe un chemin autorisé par le modèle qui suit la chronique mais qui ne comporte pas de faute.
De même, étant donnés un modèle et un ensemble de chroniques, on peut, avec des opérations classiques sur les automates, chercher un chemin qui ne suit aucune de ces chroniques et qui est d-fautive.
En combinant ces deux routines, on peut vérifier l'explainabilité du SÉD en suivant la procédure 1. Elle consiste à chercher des chemins d-fautifs et vérifier s'ils sont explicables

simplement. À chaque étape de l'algorithme, le chemin d-fautif généré doit ne pas être explicable par les chroniques déjà calculées. L'algorithme s'arrête quand il n'arrive pas à expliquer le chemin d-fautif actuel (le SÉD n'est pas explicable) ou s'il ne trouve plus de chemin d-fautif (le SÉD est explicable et on a trouvé un ensemble couvrant de chronique).
Par ailleurs, puisque le nombre de chroniques simples est borné, l'algorithme termine en un temps fini.

Procédure 1 Vérifier l'explainabilité d'un modèle

entrées : SÉD $M = \langle Q, \Sigma, T, I, \Sigma_o, F \rangle$, seuil de complexité S, délai d
sorties : `Explicable` et une liste de chroniques qui couvrent les chemins d-fautifs ou `Non explicable` et un chemin d-fautif non explicable
$\Theta = \emptyset$
boucle
    $\rho :=$ `chercher_chemin`(M, Θ, d)
    si $\rho = \bot$ **alors renvoyer** `Explicable`, Θ
    fin de si
    $\langle \theta, t \rangle :=$ `expliquer`(M, ρ, S)
    si $\theta = \bot$ **alors renvoyer** `Non explicable`, ρ
    fin de si
    $\Theta := \Theta \cup \{\theta\}$
fin de boucle

4.4 Calculer une chronique expliquant une trace

Étant donné une trace, calculer une chronique expliquant celle-ci est une tâche complexe. Nous avons déjà expliqué que toute trace peut être transformée en une chronique, mais celle-ci ne sera probablement pas simple.
À partir d'une chronique, on peut définir un espace de recherche consistant en un ensemble partiellement ordonné de chroniques. Deux chroniques $\theta_1 \preceq \theta_2$ ordonnées dans cet espace de recherche sont telles que θ_1 est une « abstraction » de θ_2, c'est-à-dire que toute séquence d'événements observables qui suit θ_1 suit également θ_2. Plus une chronique est abstraite, plus elle est simple (selon notre définition). D'autre part, on peut facilement prouver des résultats de monotonicité : si θ_2 n'est pas typique, alors θ_1 ne l'est pas non plus.
Dès lors, on peut explorer cet espace de recherche pour trouver une chronique typique simple. C'est une tâche coûteuse du point de vue calculatoire ; nos implémentations actuelles ne fonctionnent que pour des problèmes ridiculement petits. Nous proposons donc quelques pistes pour améliorer les temps de calcul :
D'une part, on souhaitera généralement partir d'une trace aussi petite que possible car l'espace de recherche sera d'autant plus petit.
Ensuite, il faudra vraissemblablement tâcher de développer des algorithmes incrémentaux : comme nous l'avons indiqué plus haut, vérifier qu'une chronique est typique requiert un produit synchrone avec le modèle. Étant donné un tel produit synchrone pour une chronique, le produit pour une autre chronique similaire comportera pour la majeure partie les mêmes états. Une procédure intelligente devrait pouvoir identifier quelles branches du produit ne nécessitent pas d'être réexplorées.
Par ailleurs, une modélisation décentralisée du système devrait permettre une étude de l'impact des observations sur le diagnostic de la faute. Les diagnostiqueurs dits dynamiques par exemple, [18, 4], identifient quels capteurs peuvent être éteints puis doivent être rallumés durant la surveillance du système. De même, si l'on est capable de déterminer qu'une observation est sans importance dans un contexte donné, on peut immédiatement l'ignorer.
D'autre part, nous prédisons que des techniques d'apprentissage automatique devraient permettre de prédire les régions de l'espace de recherche les plus intéressantes. Par exemple, on pourra chercher à générer un large nombre traces fautives et non fautives et effectuer des études statistiques pour déterminer quelles observations sont les plus fréquentes aux alentours de l'occurrence de la faute. Ces variables sont vraissemblablement impliquées dans la détection des fautes. Toute expertise extérieure pourra aider à déduire l'espace de recherche.
Enfin, nous notons que cette étape peut échouer. Ainsi, si notre algorithme est incapable de calculer une explication simple pour une trace donnée alors que celle-ci existe, on peut juste renvoyer un échec et conseiller à l'utilisateur de modifier le système pour le rendre plus explicable quand bien même ce n'est pas strictement indispensable.

5 Travaux similaires

À notre connaissance, nous sommes les premiers à proposer une définition formelle d'explication et d'explicabilité dans le cadre du diagnostic à base de modèle, même si Christopher et Grastien [5] avaient posé des jalons avec la théorie des observations critiques qui se propose d'extraire les observations importantes dans un problème de diagnostic. En comparaison, nous proposons une définition plus générique d'explication dans laquelle d'autres informations pourraient être ajoutées telle que certaines inférences intermédiaires.
Bertoglio et coauteurs [2] proposent de renvoyer une explication qui consiste à renvoyer l'ordre dans lequel les fautes sont censés avoir eu lieu, y compris si une faute a eu lieu plusieurs fois.
Des techniques de diagnostic ont été développées en IA connexionniste. Elles permettent de mettre en lumière les observations qui ont contribué le plus au diagnostic. Ainsi, Belikov et coauteurs ainsi que Brito et coauteurs [1, 3] utilisent Grad-CAM [16] tandis que Jang et coauteurs [9] se basent sur SHAP [10]. Comparées à notre travail, ces techniques manquent de garanties formelles. On peut aussi se demander comment les comportements négatifs (absence de certaines observations) peuvent être capturés par ces méthodes.
Notre approche se situe dans le cadre formel de type explication abductive [11]. Les explications y sont comprises

comme une fraction de l'entrée au processus de décision / classificateur qui garantit (quelques soient les autres entrées) la sortie effectivement observée. Comparées aux approches existantes, les entrées d'un SÉD contiennent des informations implicites sous la forme d'absence d'observation qu'il convient donc de gérer de manière spécifique.

Il existe des travaux pour l'apprentissage automatique de chroniques [12, 8, 17]. Nous pensons que les résultats présentés ici pourraient relancer cette ligne de travail.

6 Conclusion

Dans cet article, nous avons proposé la première définition d'explication de diagnostic. Une explication est un objet mathématique devant permettre à un utilisateur d'inférer le même résultat que le diagnostiqueur. Nous montrons qu'il est possible d'utiliser les chroniques pour représenter cette explication. Nous montrons également que tout système explicable peut, en première approximation, être diagnostiqué par un système de reconnaissance de chronique.

Nous considérons que ces travaux sont fondateurs pour les explications en diagnostic. Parmi les travaux futurs, il sera d'abord nécessaire de développer des techniques plus efficaces pour le calcul d'explications / chroniques. Par exemple, nous souhaitons être capable d'analyser le système pour rapidement trouver des candidats d'explication. Des techniques similaires ont été développées pour prouver la diagnosticabilité de grands systèmes [14].

Nous souhaitons également développer plus de types d'explications. Quelles informations seraient utiles pour un utilisateur ? Lorsque le diagnostiqueur mentionne telle ou telle observation, il y a peut-être une raison concrète qu'il serait utile d'expliciter. Par exemple, l'explication pourrait avoir la forme : « la lumière est allumée, ce qui prouve que la ligne est électrifiée ». Dans cet exemple, la seconde clause n'est pas informative au sens de la théorie de l'information (c'est une conséquence de la première clause), mais mentionner ce fait permet de comprendre que toute autre information qu'on pourrait déduire de la première clause peut être ignorée.

Enfin, il serait nécessaire d'étudier les critères précis qui rendent une explication simple ou qui permettent à un utilisateur de détecter une erreur de la part du diagnostiqueur. Cette question devra être résolue en conjonction avec les sciences sociales.

Références

[1] J. Belikov, M. Meas, R. Machlev, A. Kose, A. Teplja-kov, L. Loo, E. Petlenkov, and Y. Levron. Explainable AI based fault detection and diagnosis system for air handling units. pages 271–279, 2022.

[2] N. Bertoglio, G. Lamperti, M. Zanella, and X. Zhao. Explanatory monitoring of discrete-event systems. In *Intelligent Decision Technologies : Proceedings of the 12th KES International Conference on Intelligent Decision Technologies (KES-IDT 2020)*, pages 63–77. Springer, 2020.

[3] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte. Fault diagnosis using explainable AI : A transfer learning-based approach for rotating machinery exploiting augmented synthetic data. *Expert Systems with Applications*, 232 :120860, 2023.

[4] F. Cassez and S. Tripakis. Fault diagnosis with static or dynamic diagnosers. *Fundamenta Informaticae (FI)*, 88(4) :497–540, 2008.

[5] C. J. Christopher and A. Grastien. Critical observations in model-based diagnosis. *Artificial Intelligence*, 331 :104116, 2024.

[6] C. Dousson and P. Le Maigat. Chronicle recognition improvement using temporal focusing and hierarchization. In *20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 324–329, 2007.

[7] Christophe Dousson. *Suivi d'évolutions et reconnaissance de chroniques*. PhD thesis, Toulouse, 1994.

[8] B. Guerraz and C. Dousson. Chronicles construction starting from the fault model of the system to diagnose. In *Fifteenth International Workshop on Principles of Diagnosis (DX-04)*, pages 51–56, 2004.

[9] K. Jang, Karl K. Pilario, N. Lee, I. Moon, and J. Na. Explainable artificial intelligence for fault diagnosis of industrial processes. *IEEE Transactions on Industrial Informatics*, 232 :1–8, 2023.

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS-17)*, 30, 2017.

[11] J. Marques-Silva and A. Ignatiev. Delivering trustworthy AI through formal XAI. In *36th Conference on Artificial Intelligence (AAAI-22)*, 2022.

[12] E. Mayer. Inductive learning of chronicles. In *Thirteenth European Conference on Artificial Intelligence (ECAI-98)*, pages 471–472, 1998.

[13] Y. Pencolé and M.-O. Cordier. A formal framework for the decentralised diagnosis of large scale discrete event systems and its application to telecommunication networks. *Artificial Intelligence (AIJ)*, 164(1–2) :121–170, 2005.

[14] Y Pencolé. Diagnosability analysis of distributed discrete event systems. In *Fifteenth International Workshop on Principles of Diagnosis (DX-04)*, pages 173–178, 2004.

[15] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, and D. Teneketzis. Diagnosability of discrete-event systems. *IEEE Transactions on Automatic Control (TAC)*, 40(9) :1555–1575, 1995.

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM : Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV-17)*, pages 618–626, 2017.

[17] A. Subias, L. Travé-Massuyès, and E. Le Corronc. Learning chronicles signing multiple scenario instances. *IFAC Proceedings Volumes*, 47(3) :10397–10402, 2014. Nineteenth IFAC World Congress.

[18] D. Thorsley and D. Teneketzis. Active acquisition of information for diagnosis and supervisory control of discrete event systems. *Journal of Discrete Event Dynamical Systems (JDEDS)*, 17 :531–583, 2007.

[19] M. Zanella and G Lamperti. *Diagnosis of active systems*. Kluwer Academic Publishers, 2003.