La logique des incohérences : un modèle formel pour l'analyse de l'erreur humaine

Valentin Fouillard

université PARIS-SACLAY

La logique des incohérences : un modèle formel pour l'analyse de l'erreur humaine

The logic of inconsistencies: a formal model for the analysis of human error

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche **Laboratoire interdisciplinaire des sciences du numérique** (Université Paris-Saclay, CNRS) et **Laboratoire Méthodes Formelles** (Université Paris-Saclay, CNRS, ENS Paris-Saclay), sous la direction de **Nicolas SABOURET**, Professeur, le co-encadrement de **Frédéric BOULANGER**, Professeur, et le co-encadrement de **Safouan TAHA**, Maître de conférences

Thèse soutenue à Paris-Saclay, le 2 décembre 2022, par

Valentin FOUILLARD

Composition du jury

Membres du jury avec voix délibérative

Christine PAULIN-MOHRING Professeur des Universités, LMF-CNRS, Université Paris- Saclay	Présidente
Andreas HERZIG Directeur de Recherches, IRIT-CNRS, Université Paul Sabatier	Rapporteur & Examinateur
Michel OCCELLO Professeur des Universités, LCIS, Université Grenoble Alpes	Rapporteur & Examinateur
Carole ADAM Maîtresse de Conférences, LIG-CNRS, Université Grenoble-Alpes	Examinatrice
Emmanuelle GRISLIN-LE STRUGEON Maîtresse de Conférences-HdR, LAMIH-CNRS, INSA Hauts-de-France	Examinatrice

THESE DE DOCTORAT
NNT : 2022UPASG082

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : La logique des incohérences : un modèle formel pour l'analyse de l'erreur humaine
Mots clés : Diagnostic, Biais cognitifs, Révision de croyance, Logique

Résumé : Dans cette thèse, nous nous sommes intéressés à l'utilisation des méthodes formelles pour guider le diagnostic des erreurs humaines dans des situations d'accidents. L'application des méthodes formelles dans un tel contexte pose plusieurs difficultés. La première est de pouvoir expliquer à l'aide de la logique mathématique des situations incohérentes, donc en contradiction avec cette logique. La deuxième est de pouvoir comparer les différents diagnostics. En effet, une décision incorrecte n'est jamais le fruit du hasard mais se base sur les croyances, les désirs et les intentions de l'opérateur. Ainsi, toute erreur ne se vaut pas et il est nécessaire de formaliser et de définir ce qui fait un bon diagnostic.

La première partie de la thèse présente un état de l'art des travaux en sciences humaines et sociales (SHS) sur l'erreur humaine. Nous montrons qu'il est nécessaire de distinguer deux aspects : la détermination des causes d'une prise de décision erronée et la compréhension de ces causes par la recherche de biais cognitifs. Nous présentons ensuite les principaux modèles informatiques pour la modélisation du raisonnement et l'étude de l'erreur humaine. Nous montrons que le diagnostic fondé sur la cohérence (consistency-based diagnosis) et l'opérateur de révision de croyance AGM constituent une bonne piste pour l'explication d'erreurs humaines.

La deuxième partie de la thèse s'intéresse à la modélisation d'une situation d'accident et au diagnostic des décisions humaines erronées dans cette situation. Nous nous sommes basés pour cela sur une logique de croyances inspirée de la logique BDI pour la modélisation des situations d'accidents. Nous avons développé un algorithme de diagnostic itératif basé sur un opérateur de révision de croyance minimale respectant l'axiomatique AGM. Cet algorithme de diagnostic itératif a l'avantage de faciliter la distinction des erreurs de nature différentes. De plus, celui-ci est correct et complet par rapport à un algorithme de diagnostic minimal.

La troisième contribution de la thèse réside dans notre travail pour définir formellement la plausibilité d'un diagnostic. Nous nous sommes basés pour cela sur la littérature des sciences humaines et plus précisément des biais cognitifs. Pour cela, nous avons développé une première taxonomie formelle des biais qui permet de définir des caractéristiques logiques communes entre les biais. À partir de cette taxonomie, nous avons pu définir huit biais cognitifs rattachés aux biais présents dans la littérature. Nous avons ensuite considéré que plus un diagnostic peut être expliqué par les biais, plus le diagnostic est plausible.

Nous avons alors étudié la validité de ce modèle informatique sur deux cas d'étude d'accident de l'aviation civile. Nous montrons que nous retrouvons les explications proposées par le Bureau d'Enquêtes et d'Analyses ainsi que des explications non envisagées par les enquêteurs.

Nous proposons enfin plusieurs perspectives pour améliorer notre approche. Nous pensons notamment prendre en compte les émotions et les interactions sociales dans la modélisation de la situation d'accident afin d'augmenter la variété de diagnostic possible. Enfin, nous souhaitons étendre l'évaluation des diagnostics par une méta-évaluation des biais cognitifs ainsi que par la prise en compte de l'intention d'action.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: The logic of inconsistencies: a formal model for the analysis of human error
Keywords: Diagnostic, Cognitive biases, Belief revision, Logic

Abstract: In this thesis, we are interested in the use of formal methods to guide the diagnosis of human errors in accident situations. The application of formal methods in such a context raises several difficulties. The first one is to be able to explain with the help of mathematical logic situations that are incoherent and therefore in contradiction with this logic. The second is to be able to compare the different diagnoses. Indeed, an incorrect decision is never the work of a hazard but is based on the beliefs, desires and intentions of the operator. Thus, not all errors are equal and it is necessary to formalize and define what makes a good diagnosis.

The first part of the thesis presents a state of the art of human and social sciences (HSS) work on human error. We show that it is necessary to distinguish two aspects : the determination of the causes of erroneous decision making and the understanding of these causes through the search for cognitive biases. We then present the main computer models for modeling reasoning and studying human error. We show that consistency-based diagnosis and the belief revision operator AGM is a good way to explain human errors.

The second part of the thesis deals with the modeling of an accident situation and the diagnosis of human errors in this situation. We have based our work on a belief logic inspired by the BDI logic for the modeling of accident situations. We have developed an iterative diagnosis algorithm based on a minimal belief revision operator respecting the AGM axiomatic. This iterative diagnosis algorithm has the advantage of facilitating the distinction of errors of different nature. Moreover, it is correct and complete compared to a minimal diagnosis algorithm.

The third contribution of the thesis lies in our work to formally define the plausibility of a diagnosis. We based our work on the literature of human sciences and more precisely on cognitive biases. For this purpose, we have developed a first formal taxonomy of biases that allows us to define common logical characteristics between biases. From this taxonomy, we were able to define eight cognitive biases related to the biases present in the literature. We then considered that the more a diagnosis can be explained by the biases, the more plausible the diagnosis is.

We then studied the validity of this computer model on two cases of civil aviation accidents. We show that we find the explanations proposed by the Bureau d'Enquêtes et d'Analyses as well as explanations not considered by the investigators.

Finally, we propose several perspectives to improve our approach. In particular, we intend to take into account emotions and social interactions in the modeling of the accident situation in order to increase the variety of possible diagnoses. Finally, we wish to extend the evaluation of the diagnoses by a meta-evaluation of the cognitive biases as well as by taking into account the intention of action.

Remerciements

Tout d'abord, je tiens à remercier mon directeur de thèse Nicolas Sabouret pour ses encouragements, sa confiance, son optimisme à toute épreuve, sa rigueur pour synthétiser ma pensée et d'avoir toujours trouvé du temps pour moi. Je ne pouvais pas espérer mieux comme directeur. J'ai adoré collaborer avec toi et je souhaite sincèrement pouvoir dans le futur réitérer l'expérience. Je souhaite aussi remercier Safouan Taha, mon encadrant de thèse qui rechargeait mes batteries avec nos sessions de travail en duo, nos discussions informelles et son enthousiasme face à mes travaux. Ta rigueur et tes doutes sur l'aspect formel de ma thèse m'ont permis d'améliorer grandement mes compétences dans ce domaine. Je remercie aussi Frédéric Boulanger, mon second encadrant qui par ses conseils, ses relectures et ses cours de Latex m'a toujours permis d'aller plus loin dans la pédagogie et la présentation de mes idées. Je vous remercie tous les trois pour cet accompagnement sur les trois ans.

Mes remerciements vont aussi à mon jury de thèse. Merci à Andreas Herzig et Michel Occello, pour leurs échanges et retours constructifs. Merci à Christine Paulin-Mohring et Carole Adam pour l'intérêt qu'elles portent à mon travail tout comme Emmanuelle Grislin-Le Strugeon que je remercie aussi pour ses conseils et retours à mi-parcours.

Cette thèse ne serait pas la même sans l'ambiance plaisante entre les membres de l'équipe CPU. Je tiens donc à remercier Jean-Baptiste, Amine, Morghane, David, Delphine, Rachel, Brian, Émilie et tous les autres pour leur aide ou discussions autour du café. Plus particulièrement, je tiens à remercier un trio sans lequel cette thèse n'aurait pas eu la même saveur. Jennifer, toujours présente pour aider, toujours les bons mots au bon moment ou pour changer les idées, tu as rendu cette thèse beaucoup plus douce et moins amère à vivre et mérites mille mercis. Florian, ton non-sérieux (c'est un compliment !), ta bonne humeur quotidienne et ton humour n'ont rendu que plus agréables et courtes les journées de travail. Grâce à toi, la veille ne ressemblait pas au lendemain et pour toi ce n'est peut-être rien, mais pour moi c'est beaucoup ! Séverine qui a su souder les doctorants à la sortie d'une période difficile par sa bonne humeur contagieuse, ses conseils, sa cuisine, sa philosophie de vie et de travail (très) inspirante, tu es une des raisons majeures de l'ambiance agréable qui règne entre doctorants aujourd'hui. Ne changez surtout pas !

D'un point de vue plus personnel, je tiens bien sûr à remercier mon père, ma mère, ma petite soeur, mon petit frère et ma famille pour leur soutien quotidien sans faille et toutes les autres choses qui ne tiendraient pas sur cette page. Je remercie aussi mes gars sûrs Benoit, Florent, Pierre et Julie toujours là pour changer les idées presque tous les soirs et tous les week-ends. On se connaît depuis bien trop longtemps pour déballer tout le bien que je pense de vous, vous le savez.

Enfin, je remercie infiniment la personne qui me connaît par coeur, sait ce que ce travail représente pour moi et qui a été là à chaque instant des plus heureux au moins heureux :

عزيزة، شكرا جزيلا على كل شيء، أحبك

Table des matières

I - Introduction

I.1 Contexte général et présentation de la problématique

Une décision humaine n'est pas toujours logique. N'avez vous jamais pris la décision de préférer le restaurant ayant le plus de monde qu'un restaurant vide ? De considérer que si la pièce tombe sur pile plusieurs fois d'affilé alors elle a plus de chance de tomber sur face au coup suivant ? Pourtant un restaurant vide n'est pas synonyme de mauvais restaurant, tout comme un restaurant bondé n'est pas synonyme d'un bon restaurant. De plus, la probabilité que la pièce tombe sur pile ou sur face est complètement indépendante du précédent tirage. Ces exemples de comportements illogiques ont déjà été étudiés dans la littérature sous le terme de *Bandwagon effect* pour l'exemple du restaurant [Schmitt-Beck, 2015] et de *Gambler's fallacy* [Tversky et al., 1971] pour l'exemple de la pièce. Bien que sur ces exemples la décision illogique est sans grande conséquence, dans d'autres situations de telles décisions peuvent avoir de lourdes conséquences. C'est le cas par exemple dans l'aviation ou la médecine où une décision erronée peut mener à des accidents et à la mort d'un ou plusieurs individus [Murata et al., 2015]. Il est donc important de comprendre pourquoi ces décisions illogiques sont prises par des humains afin d'éviter que des situations similaires puissent se répéter. Dans ce sens, lors d'un accident, une investigation est mise en place par des enquêteurs dont l'objectif est de déterminer les raisons de l'incident. C'est le cas par exemple des enquêteurs du Bureau d'Enquêtes et d'Analyses (BEA) pour la sécurité de l'aviation civile qui sont spécialisés dans les accidents d'aviation. Dans cette thèse, nous nous intéressons à l'utilisation des modèles informatiques pour faciliter cette investigation. Nous pensons que les méthodes formelles sont des outils adaptés pour répondre à cet enjeu.

En effet, ces méthodes permettent de *spécifier* un système, c'est-à-dire à partir d'un langage mathématique logique, définir la description d'un système et ses propriétés. Elles permettent aussi de *vérifier* une propriété d'un système, c'est-à-dire s'assurer que cette propriété est vraie dans le système [Woodcock et al., 2009]. Dans notre cas, l'application des méthodes formelles consiste à *spécifier* la situation

de l'accident et *vérifier* qu'une hypothèse d'une erreur humaine permet d'expliquer l'accident. Les méthodes formelles auraient l'avantage de :

— Assurer la pertinence de l'hypothèse d'une erreur humaine car s'il manque des informations pour vérifier cette propriété sur l'accident spécifié alors l'hypothèse sera rejetée. Les méthodes formelles obligeraient les enquêteurs à n'oublier aucune information dans la spécification de l'accident et la validation de l'hypothèse.

— Explorer des hypothèses de diagnostic auxquelles les enquêteurs n'auraient pas pensé. En effet, si une erreur humaine peut-être définie comme une propriété d'un système d'accident alors il est possible grâce aux méthodes formelles d'explorer toutes les solutions vérifiant cette propriété.

— Définir formellement la plausibilité d'un diagnostic, c'est-à-dire spécifier les critères utilisés pour comparer les différentes hypothèses de diagnostic.

Ainsi, nous pensons que les méthodes formelles permettraient d'imposer une définition rigoureuse de la situation de l'accident, des hypothèses, de la plausibilité des hypothèses, de vérifier formellement les hypothèses des erreurs humaines et d'explorer potentiellement des hypothèses d'explications non explorées par les enquêteurs.

L'utilisation de méthodes formelles pose néanmoins de nombreuses difficultés dans le contexte de l'erreur humaine. En effet, ces méthodes se basant sur un langage logique, les propriétés vérifiées sur le système spécifié doivent être logiquement déduites du système. Or l'erreur humaine n'a rien de logique, car par définition celle-ci diffère d'un comportement attendu. Appliquer les méthodes formelles dans ce contexte revient donc à vérifier une propriété illogique avec des méthodes logiques. Une première question alors émerge :

Peut-on à partir d'un langage logique vérifier des propriétés illogiques ?

Plus précisément, dans notre contexte :

Peut-on à partir d'un langage logique déduire des comportements illogiques d'un humain ?

Au-delà de cette première interrogation, une autre difficulté émerge : vérifier des propriétés illogiques est différent de vérifier des propriétés logiques. Dans le cas où une propriété est déduite logiquement, aucune question ne se pose sur la pertinence de la propriété : elle résulte naturellement du système spécifié par le langage logique. Ce n'est pas le cas pour une propriété illogique. En effet, l'erreur humaine résulte d'un raisonnement qui n'est pas le fruit du hasard. Celui-ci se base sur les croyances de l'agent, ses désirs, ses intentions, etc. Par conséquent, toute erreur ne se vaut pas, cela veut dire que toute propriété illogique ne se vaut pas. Une deuxième problématique de cette thèse est alors :

Comment vérifier la pertinence d'une propriété illogique ?

Plus précisément, dans notre contexte :

Comment vérifier qu'un comportement illogique est plus pertinent qu'un autre dans le contexte d'un accident ?

Ces travaux de thèse se situent donc au croisement de deux disciplines : l'informatique et les sciences humaines. Les sciences humaines pour l'étude de l'erreur humaine et l'informatique pour les méthodes formelles.

I.1.1 Objectifs

L'enjeu d'utiliser les méthodes formelles pour le diagnostic des erreurs humaines dans une situation d'accident afin de bénéficier des avantages de ces méthodes, nous a conduit à établir la problématique suivante :

Diagnostiquer les prises de décisions erronées dans une situation d'accident à l'aide de méthodes formelles

De cette problématique, et des différentes questions de recherche abordées précédemment, découlent les objectifs suivants :

- Proposer un langage de modélisation pour spécifier une situation d'accident.
- Déterminer formellement si une erreur humaine est pertinente pour expliquer une décision erronée dans la situation d'accident.
- Proposer un algorithme de diagnostic permettant de générer des explications pour des décision erronées.
- Déterminer formellement la plausibilité d'un diagnostic et par conséquent la ou les meilleurs diagnostics d'erreur humaine pour expliquer une décision erronée.

I.2 Organisation du manuscrit

Ce manuscrit est organisé en sept chapitres. Le présent chapitre étant considéré comme une introduction globale à la problématique de recherche.

Le chapitre II est un état de l'art de l'erreur humaine dans les sciences humaines où nous nous intéressons dans un premier temps comment est étudié l'erreur humaine dans le domaine de la sécurité. Dans un deuxième temps, nous nous intéressons à l'étude de l'erreur humaine en science cognitive à travers les biais cognitifs. Plus précisément, quelles sont les différentes taxonomies présentes dans la littérature permettant de caractériser ces biais.

Le chapitre III est un état de l'art des méthodes et modèles informatiques pour la modélisation du raisonnement et l'étude de l'erreur humaine. Dans un premier temps nous nous intéressons aux logiques de croyances et d'actions et aux

difficultés qui doivent être prises en compte lors de la modélisation d'un agent rationnel dans un monde dynamique. Dans un deuxième temps, nous abordons les différentes approches possibles pour diagnostiquer un système formel ainsi que la relation qui existe entre certaines de ces approches et la révision de croyance. Dans un troisième temps, nous présentons les différents modèles formels permettant de capturer des biais cognitifs. Enfin nous nous positionnons sur les théories et travaux présentés sur les deux états de l'art.

Le chapitre IV est une définition informelle de l'approche adoptée dans ces travaux. Dans un premier temps, nous introduisons l'accident du vol Rio-Paris qui servira de cas d'étude dans ce manuscrit. Dans un deuxième temps, nous illustrons notre approche à partir de ce cas d'étude en définissant le modèle d'*explication* chargé de déterminer les états de croyances cohérents de l'agent et le modèle d'*évaluation* qui attribue une plausibilité pour chaque état de croyances.

Le chapitre V définit formellement le diagnostic des états de croyances de l'agent. Dans un premier temps, nous abordons la syntaxe et les éléments du modèle. Nous donnons ensuite la sémantique des états de croyances de l'agent. Dans un troisième temps nous définissons le principe de recherche de cohérence pour notre opérateur de diagnostic. A la suite de cette section, nous montrons l'équivalence de notre opérateur de diagnostic avec un opérateur AGM. Enfin nous appliquons cette opérateur sur le cas d'étude de l'accident Rio-Paris et discutons des résultats.

Le chapitre VI s'intéresse à la prise en compte de l'*inertie* des croyances dans notre modèle. Pour illustrer la problématique, nous proposons dans un premier temps une formalisation de l'accident du mont Sainte-Odile. Nous définissons ensuite la problématique de la non-conservation des croyances d'un pas de temps au suivant, que nous appelons « distortion du décors » par analogie avec le problème du décor de [McCarthy et al., 1969]. Nous proposons ensuite une solution pour prendre en compte l'*inertie* dans notre modèle. De plus nous discutons des nouvelles incohérences introduites par l'*inertie* des croyances et comment notre opérateur de diagnostic peut prendre en compte ces nouvelles incohérences. Nous proposons à partir de là un algorithme de diagnostic par itération basé sur notre opérateur de diagnostic afin de permettre la distinction des natures des ignorances et faciliter la compréhension du diagnostic. Nous définissons à partir de ces itérations la notion de choix pour l'agent puis montrons la correction et complétude de l'algorithme de diagnostic par itération.

Le chapitre VII s'intéresse au modèle d'*évaluation* qui détermine la plausibilité des états de croyances trouvés par le modèle d'*explication*. Nous définissons dans un premier temps une première taxonomie de biais cognitifs basée sur les ignorances, croyances et choix de l'agent. Nous définissons ensuite huit biais avec les caractéristiques présentes dans la taxonomie. Dans un troisième temps, nous définissons un algorithme de filtrage des états de croyances peu plausibles en fonction des biais cognitifs permettant d'expliquer les ignorances des états de croyances. Enfin nous

appliquons cet algorithme de filtrage sur les deux cas d'études et discutons des résultats.

Enfin ce manuscrit s'achève sur une conclusion présentant une vue globale de notre approche ainsi que les différentes contributions de nos travaux dans le domaine de l'intelligence artificielle. Nous détaillons également les perspectives nouvelles que soulèvent ces travaux.

II - L'erreur humaine dans les sciences humaines

Errare humanum est

Sénèque

L'humain est un être non-omniscient et imparfait, ce qui a pour effet que celui-ci commet parfois des erreurs. Ces erreurs n'ont souvent aucune conséquence, mais peuvent parfois avoir des résultats désastreux lorsque l'humain opère dans un système complexe [Murata et al., 2015]. Il y a donc sur la question de l'erreur humaine à la fois un enjeu scientifique sur la compréhension du raisonnement humain et à la fois un enjeu industriel et sécuritaire afin d'éviter des pertes économiques ou pire humaines sur les conséquences de ces erreurs. C'est pourquoi, la littérature scientifique s'est intéressée à l'étude de l'erreur humaine, notamment à travers :

— le domaine de la sécurité qui s'intéresse à trouver les facteurs humains qui permettent d'expliquer les accidents [Wiegmann et al., 2017] ;
— le domaine des sciences cognitives qui s'intéresse à trouver les processus cognitifs qui rentrent en jeu lorsqu'un agent prend une décision erronée [Tversky et al., 1974].

Nous allons dans ce chapitre voir comment ces deux domaines étudient la question de l'erreur humaine. Plus précisément, nous aborderons comment, dans le domaine de la sécurité, les enquêteurs sur les accidents raisonnent pour déterminer l'erreur humaine. Nous verrons ensuite que les travaux en sécurité et sciences cognitives sont complémentaires. Enfin, nous aborderons l'étude et la classification des erreurs humaines en sciences cognitives.

II.1 Erreur humaine en situation d'accident

II.1.1 La multicausalité des accidents

Enquêter sur un accident est complexe, car cela nécessite de prendre en compte une multitude de facteurs relevant de plusieurs couches. L'erreur peut provenir de la décision d'un humain, de problèmes de communication dans la hiérarchie du système étudié, d'une panne technique d'un appareil, etc. Non seulement les raisons d'un accident peuvent être de natures différentes, mais aussi multiples. Par exemple, une panne technique peut entraîner un problème de communication entre les opérateurs du système et, par effet boule de neige, entraîner une erreur de décision. C'est pourquoi la littérature dans le domaine de la sécurité s'intéresse à représenter le fonctionnement d'une situation d'accident.

Le modèle le plus connu de la littérature est le modèle du fromage suisse de [Reason, 1990] (Voir figure II.1) où chaque couche de fromage correspond à un acteur du système étudié, c'est-à-dire un outil ou un opérateur qui joue un rôle dans les décisions prises sur le système. Cela peut être par exemple dans le cas de l'aviation, le pilote, la tour de contrôle, les outils de navigation, etc. Chaque couche de fromage comporte des trous qui représentent des failles sur un acteur du système. Un accident survient lorsque tous les trous sont alignés. Concrètement, le modèle de Reason dit qu'un accident ne peut être résumé à une erreur, mais une succession d'erreurs entre les différents acteurs du système. Le but de la sécurité est alors de réduire au maximum les trous de chaque couche afin que les différents acteurs ne laissent pas passer l'erreur de la couche précédente.

Dans ce chapitre, nous nous intéressons à l'erreur humaine, c'est-à-dire à une unique couche du modèle de Reason qui correspond à un humain. Nous allons aborder dans la suite comment le domaine de la sécurité recherche ces failles empruntées par la couche de l'opérateur humain.

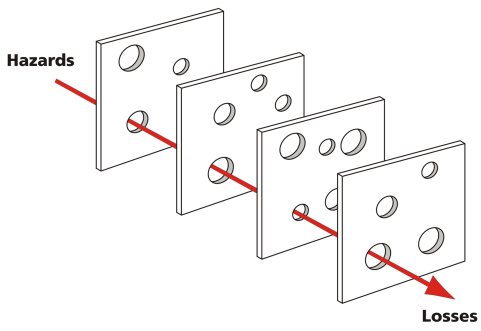


Figure II.1 – Modéle du fromage suisse de Reason
(source : Davidmack, CC BY-SA 3.0)

II.1.2 Trouver et comprendre l'erreur humaine

Dans son livre *The field guide to understanding 'human error'*, [Dekker, 2006] développe l'idée du fromage suisse de Reason et considère l'humain comme une couche de fromage. Il oppose alors deux visions de l'erreur humaine. La première nommée "l'ancienne vision" qui considère l'erreur humaine comme une cause de l'incident et la deuxième nommée "la nouvelle vision" qui considère l'erreur humaine comme le symptôme d'incidents plus profonds dans le système où l'humain est un acteur. Ces deux paradigmes s'opposent aussi dans la recherche de l'explication de l'erreur. Dekker résume ces différences par le tableau suivant :

Ancienne vision	**Nouvelle vision**
L'erreur humaine est la cause de l'incident	L'erreur humaine est le symptôme d'incidents plus profonds dans le système
Pour expliquer l'échec, vous devez rechercher l'échec	Pour expliquer un échec, n'essayez pas de trouver où les gens se sont trompés
Vous devez trouver les évaluations inexactes, les mauvaises décisions et les mauvais jugements	Trouvez plutôt comment les évaluations et les actions avaient un sens à l'époque compte tenu des circonstances qui les entouraient

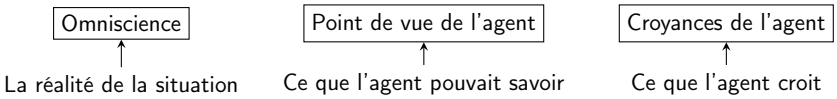
Dekker met en avant un point important dans son argumentation. La notion de décision incorrecte, c'est-à-dire une décision qui ne devait pas être faite, a peu de sens : "*In the sequence of events leading up to failures, there is no 'badness' in anybody's behavior by any objective measure.*" Dekker considère que l'humain agit rationnellement mais de manière locale (*i.e* avec ses connaissances et les informations à disposition) : "*The point is, people in safety-critical jobs are very likely doing the right thing under the circumstances. They are doing reasonable things given their point of view and focus of attention; their limited knowledge of the situation; their objectives.*" De ce fait, dire qu'une décision est incorrecte est une erreur pour Dekker car l'humain cherche la meilleure décision qui a du sens pour lui. Si nous revenons sur le modèle du fromage suisse, un opérateur humain peut recevoir par exemple des informations erronées de la couche précédente et donc prendre la meilleure décision de son point de vue avec ces informations.

Pour comprendre l'erreur humaine dans un accident, la méthode de Dekker consiste donc à déterminer les croyances de l'agent qui sont cohérentes avec la décision erronée. L'auteur met en avant tout au long de son livre l'importance d'utiliser ce nouveau paradigme afin d'éviter :

— de tomber dans le biais rétrospectif, c'est-à-dire surestimer que l'incident aurait pu être évité ;
— de considérer que seul l'humain est responsable de l'incident ;
— de considérer qu'il existe une cause fondamentale de l'incident.

II.1.3 Trois niveaux d'analyse

À partir des travaux de Dekker, nous proposons de définir trois niveaux d'analyse pour un accident :

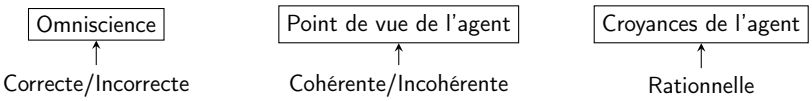


(1) L'omniscience qui correspond à "l'ancienne vision" de Dekker : la décision de l'agent est jugée à la connaissance de toutes les informations disponibles sans considérer celles que l'agent n'avait pas à sa disposition.
(2) Le point de vue de l'agent : la décision de l'agent est jugée en fonction des informations à la disposition de l'agent et du contexte dans lequel se trouve l'agent.
(3) Les croyances de l'agent : la décision de l'agent est jugée en fonction de ce que l'agent croyait dans la situation et de comment celui-ci pouvait interpréter les différentes informations à sa disposition.

L'objectif d'un enquêteur pour Dekker est donc d'analyser l'accident au niveau des croyances (3). Toute la difficulté étant qu'en tant qu'enquêteur, il analyse la situation après coup et se retrouve alors par défaut au niveau de l'omniscience (1). De ces trois niveaux d'analyse, nous allons proposer dans la section suivante trois notions de décisions qui en découlent.

II.1.4 Trois notions de décision

Au vu des trois niveaux d'analyse définit précédemment, nous proposons qu'une décision puisse être "mauvaise" ou "bonne" sur trois niveaux. Considérons la décision d'un pilote en décrochage pour illustrer notre propos. Lorsque qu'un avion est en décrochage, la "bonne" décision à prendre est de pousser le manche de contrôle. La "mauvaise" décision est de tirer le manche de contrôle. Ainsi, une décision peut être :



(1) — **Correcte**, c'est-à-dire que la décision est attendue par rapport à la réalité du monde. Par exemple, le pilote pousse le manche lorsque l'appareil est en décrochage.

— **Incorrecte**, c'est-à-dire que la décision n'est pas attendue par rapport à une réalité du monde et un objectif donné. Par exemple, le pilote tire sur le manche alors que l'appareil est en décrochage.

(2) — **Cohérente**, c'est-à-dire que la décision est attendue par rapport aux informations qui sont disponibles pour l'agent. Par exemple, l'agent peut entendre l'alarme de décrochage et décider de pousser le manche de contrôle.

— **Incohérente**, c'est-à-dire que la décision n'est pas attendue par rapport aux informations qui sont disponibles pour l'agent. Par exemple, l'agent peut entendre l'alarme de décrochage et décider de tirer le manche de contrôle.

(3) — **Rationnelle**, c'est-à-dire que la décision est attendue par rapport aux croyances de l'agent. Par exemple, l'agent croit qu'il est en décrochage et pousse sur le manche.

— **Irrationnelle**, c'est-à-dire que la décision est inattendue par rapport aux croyances de l'agent. Par exemple, l'agent croit qu'il est en décrochage, mais décide de tirer le manche de contrôle.

Pour Dekker, la notion de décision irrationnelle ne peut exister, un agent humain est toujours rationnel et prend toujours la meilleure décision avec ses croyances. De ce fait, dans le paradigme de Dekker, il existe quatre possibilités :

(a) *Une décision correcte, cohérente et rationnelle* est le meilleur des cas, l'agent prend la bonne décision avec les informations à sa disposition.

(b) *Une décision correcte, incohérente et rationnelle* est le cas où l'agent prend une bonne décision malgré lui. C'est-à-dire qu'il prend une décision qui n'est pas cohérente avec les informations à sa disposition, mais s'avère être la bonne. Par exemple, l'agent décide de pousser le manche de contrôle alors que l'avion est en décrochage, alors que les instruments de pilotage lui indiquent de tirer le manche.

(c) *Une décision incorrecte, cohérente et rationnelle* est typiquement le cas où l'agent reçoit des informations erronées qui l'amènent à prendre une mauvaise décision.

(d) *Une décision incorrecte, incohérente et rationnelle* est le cas où l'agent prend une mauvaise décision même au vu des informations perçues. C'est-à-dire que l'agent fait un raisonnement ou interprète de manière erronée les informations perçues. C'est le cas dans l'accident de Rio-Paris, le pilote peut observer l'alarme de décrochage, mais décide de tirer le manche de contrôle pendant un décrochage, car il croit être en survitesse.

Le paradigme de Dekker montre donc l'importance de différencier ces trois notions de décision pour expliquer l'erreur humaine dans une situation d'accident afin de comprendre une décision "mauvaise" que ce soit sur une analyse omnisciente ou du point de vue de l'agent. Plus précisément, trois types de décision erronée sont à considérer et correspondent aux décisions (b), (c) et (d) décrites ci-dessus

(la (a) étant une décision parfaitement correcte).

II.1.5 Conclusion

Comprendre l'erreur humaine dans une situation d'accident est un problème difficile, car l'erreur s'inscrit dans un système où l'erreur humaine peut résulter d'un enchaînement complexe d'erreurs de plusieurs acteurs dans la situation. La littérature de la sécurité propose donc, afin de ne pas considérer que l'humain est le seul responsable de l'accident et afin d'éviter tout biais rétrospectif, de déterminer les croyances de l'agent humain dans la situation. Nous faisons découler de cette proposition trois niveaux d'analyse et trois notions de décision rattachées à chaque niveau. Les deux niveaux d'analyse importants dans la littérature de la sécurité sont : le point de vue de l'agent qui correspond au contexte et aux informations à la disposition de l'agent à un instant donné et les croyances de l'agent qui sont les informations que l'agent croit et interprète à un instant donné. Une décision est alors dite incohérente si elle n'est pas attendue du point de vue de l'agent. Enfin, le modèle de Dekker fait l'hypothèse qu'un agent humain prend toujours une décision qui est rationnelle par rapport à ses croyances. Ainsi, trouver les croyances d'un agent revient à comprendre la décision de celui-ci. De ce fait :

> **La clé pour diagnostiquer une décision incorrecte est donc de se placer au niveau d'une décision incohérente et de déterminer comment elle était rationnelle avec les croyances de l'agent.**

Nous verrons toutefois dans la section suivante que déterminer les croyances d'un agent ne permet pas un diagnostic complet pour comprendre la décision de l'agent et que d'autres outils sont nécessaires pour améliorer le diagnostic.

II.2 Le pourquoi du comment

Nous avons vu dans les sections précédentes que pour le domaine de la sécurité, déterminer les croyances de l'agent pour comprendre un accident est essentiel. La méthode de Dekker nous donne une explication à une décision erronée de type :

L'agent a pris la mauvaise décision d car il croyait φ

Dans l'exemple de Rio-Paris, le pilote a pris la mauvaise décision de tirer le manche de contrôle, car il croyait qu'il était en survitesse. Une telle explication nous renseigne sur le **"comment"** de la décision qui a été prise. C'est-à-dire les causes de la prise de décision de l'agent : c'est le fait de croire en φ qui est la cause de la décision erronée. Les croyances sont donc une justification de la décision prise par l'agent. Toutefois, cela ne permet pas de complètement expliquer la prise de décision de l'agent.

En effet, tout comme la décision, les croyances doivent avoir une justification afin de comprendre l'ensemble du raisonnement de l'agent :

L'agent a pris la mauvaise décision d car il croyait φ **mais pourquoi ?**

Sans cette justification des processus qui ont amené les agents à une certaine croyance, les systèmes où agissent ces agents seront condamnés à pousser aux mêmes erreurs du fait que le raisonnement sous-jacent de l'erreur de décision n'est pas compris [Murata et al., 2015, Dekker, 2006]. Il ne faut donc pas seulement le "comment" mais aussi le **"pourquoi du comment"** pour trouver les raisons d'une croyance d'un agent.

Les rapports d'accidents montrent que les enquêteurs suivent ce principe en cherchant à la fois le *comment* et le *pourquoi*. Par exemple, dans le rapport de l'accident de Rio-Paris [BEA, 2012], nous pouvons trouver que les enquêteurs cherchent une explication à la décision de « la position du manche maintenue arrière ou neutre [qui] a continué à aggraver la situation. »(p.185). Cette décision est justifiée parce que :

— *Comment ?* : « il est possible que le [pilote] ait identifié une situation de survitesse. »(p.186)
— *Pourquoi ?* : « il est possible qu'un phénomène de sélectivité attentionnelle ait réduit sa capacité de perception de l'alarme [de décrochage]. » (p.188)

Rechercher le pourquoi est donc essentiel dans un diagnostic complet d'une erreur humaine. Nous verrons dans la section suivante comment le domaine des sciences cognitives, qui s'intéressent aux processus cognitifs dans le raisonnement humain, étudie et définit les différentes erreurs possibles.

II.3 L'erreur humaine en sciences cognitives

II.3.1 Définition des biais cognitifs

Les sciences cognitives étudient les erreurs humaines à travers les limitations cognitives humaines et plus précisément les biais cognitifs. L'humain utilise des raccourcis de pensée (heuristiques) [Tversky et al., 1974] pour compenser sa *rationalité limitée* [Simon, 1990], c'est-à-dire ses limitations en termes de mémoire et de temps. Par exemple, une heuristique dite de *availability* consiste à évaluer la probabilité d'un évènement en fonction de sa disponibilité dans la mémoire.

Ces heuristiques peuvent être très efficaces. [Gigerenzer et al., 2009] montre en effet que l'utilisation de l'heuristique *Take-the-Best* est parfois plus efficace qu'une technique de régression linéaire multiple. Cette heuristique consiste, lorsqu'il y a un choix entre deux options, à rechercher la première caractéristique qui discrimine ces deux options et à choisir l'option qui a la meilleure valeur sur cette caractéristique. Toutefois, ces heuristiques ne sont pas parfaites par définition et peuvent mener à des erreurs. Une erreur qui résulte d'une heuristique est nommée **biais cognitif**.

II.3.2 Les taxonomies de biais

À la suite des travaux fondateurs de Tversky et Kahneman, une littérature florissante est née sur l'étude des biais cognitifs. [Dimara et al., 2020] dénombre 151 biais cognitifs dans la littérature scientifique, présents dans de nombreux domaines, avec des erreurs produites par ces biais qui ont parfois des conséquences lourdes. Par exemple [Murata et al., 2015] analyse 5 accidents dont le crash de Tenerife [Board, 1979], le crash de la navette Challenger [Commission, 1986], l'accident nucléaire de Three Mile Island [Commission, 1979] et met en avant le rôle de plusieurs biais cognitifs. Par exemple, le biais de confirmation (*i.e.* la tendance à rechercher des informations en accord avec nos croyances et à rejeter les informations contradictoires [Nickerson, 1998]) et le biais d'optimisme (*i.e* la tendance à surestimer les évènements positifs et sous-estimer les évènements négatifs [Sharot, 2011]). D'autres travaux montrent la présence non négligeable de ces biais dans des situations d'accident comme dans le domaine nucléaire [Takano et al., 1999], sur la conduite dangereuse [Măirean et al., 2021] ou encore chez les pilotes d'avion [Walmsley et al., 2016].

Du fait du nombre important de ces biais et de leurs différents domaines d'application, une partie de la littérature s'est concentrée sur les caractéristiques communes de ces biais en développant différentes taxonomies. Ces taxonomies présentent un intérêt tout particulier dans notre approche, car elles permettent de se rapprocher d'une définition plus formelle des biais en les regroupant selon des caractéristiques précises.

Pour classifier les différents biais, plusieurs méthodes ont été proposées dans la littérature. Nous allons dans cette partie nous baser sur les états de l'art de [Dimara et al., 2020] et [Ceschi et al., 2019] qui offrent à eux deux un horizon très complet des approches et problématiques de ces taxonomies. À partir de ces états de l'art, les taxonomies de biais peuvent être séparées en trois approches :

- Les approches fondées sur les modèles qui cherchent à classer les biais selon le processus à l'origine des biais (généralement une heuristique).
- Les approches basées sur les tâches qui cherchent à classer les biais selon le type de tâche effectuée.
- Les approches empiriques qui cherchent à classer les biais par la mesure des différences individuelles lors d'une tâche de prise de décision.

Nous allons pour chacune de ces approches présenter des exemples ainsi que les avantages et limitations de celles-ci.

II.3.2.a Taxonomies fondées sur les modèles

Modèles d'heuristiques Les premières taxonomies fondées sur les modèles dans la littérature se basent sur l'idée que les biais sont les résultats d'une heuristique et que, de ce fait, les biais peuvent être classés en fonction de celles-ci. Le but de ces approches est alors de trouver le nombre d'heuristiques minimal qui per-

met d'expliquer les biais et d'attribuer chaque biais à l'heuristique correspondante. Dans leurs travaux fondateurs [Tversky et al., 1974] définissent trois heuristiques possibles :

- *Availability* : la tendance à évaluer la probabilité d'un événement en fonction du degré de disponibilité de celui-ci en mémoire.
- *Representativeness* : la tendance à évaluer la probabilité d'un évènement en fonction de la similarité avec un stéréotype.
- *Anchoring and Adjustement* : la tendance d'ajuster une valeur à partir d'une valeur initiale.

Toutefois, ces trois heuristiques ne suffisent pas à elles seules à expliquer tous les biais. Les biais sociaux ne sont par exemple pas pris en compte comme la *soumission à l'autorité*, qui est la tendance des personnes à ne pas remettre en cause une figure d'autorité au point d'effectuer des actes contre leurs principes moraux [Milgram, 1963]. C'est pourquoi, à la suite de ces travaux, d'autres modèles ont ajouté de nouvelles heuristiques [Kahneman et al., 1982] supplémentaires à considérer pour classer les biais, par exemple des heuristiques sociales [Plous, 1993], des heuristiques liées à des préférences [Camerer et al., 2004] (*i.e* le choix entre plusieurs possibilités) ou des heuristiques affectives [Finucane et al., 2000].

Modèles cognitif Une autre manière d'expliquer et de classer les biais est d'utiliser un modèle cognitif du raisonnement humain. L'un des plus utilisés dans la littérature de la taxonomie des biais est le *dual-process model* [Frankish, 2010] qui consiste à différencier deux types de raisonnements indépendants appelés *System 1* et *System 2*. Le *System 1* permet de prendre une décision rapide et efficace, généralement de manière inconsciente, là où le *System 2* permet de prendre une décision plus lente et coûteuse, mais analytique. En se basant sur ce modèle, [Kahneman, 2003] classe les biais en fonction de ces deux processus. Le *System 1* comporte par exemple des biais qui se rattachent à l'émotion (*e.g* l'excès de confiance, la tendance à considérer son jugement comme bien plus précis et performant qu'il ne l'est vraiment) là où le *System 2* comporte par exemple des biais liés à une erreur statistique (*e.g* l'oubli de la fréquence de base qui est la tendance à négliger la taille de l'échantillon d'un événement lorsqu'on calcule sa fréquence). Enfin, nous pouvons citer [Stanovich, 2009] qui n'utilise pas un *dual-process model* mais un *Tripartite model* où le *System 2* est divisé en deux processus nommés *Reflective* et *Algorithmic*. Le premier processus concerne la formation des connaissances, croyances et buts de l'agent comme le *biais rétrospectif* qui consiste à percevoir les évènements du passé comme prédictibles. Le deuxième processus concerne les stratégies cognitives adoptées comme l'*erreur du parieur* qui consiste à considérer par exemple que lorsque la pièce tombe plusieurs fois à la suite sur pile, elle a plus de chance de tomber sur face au prochain coup.

Il existe donc de nombreuses taxonomies des biais cognitifs fondées sur les modèles. Néanmoins, ces taxonomies ne font pas une liste de tous les biais présents dans la littérature pour les organiser, mais offrent un modèle abstrait pour expliquer

plusieurs biais à travers quelques exemples. De ce fait, il est possible que certains biais ne soient pas pris en compte dans les différents modèles, ce qui explique l'ajout de nouvelles heuristiques dans les modèles d'heuristique ou de processus dans les modèles cognitifs. Enfin, [Gigerenzer et al., 2009] montre que les concepts décrits dans ces modèles sont flous, imprécis et difficiles à réfuter. Ce dernier point met en avant la nécessité d'une approche plus formelle afin de définir précisément les caractéristiques communes et différentes des biais.

II.3.2.b Taxonomies fondées sur les tâches

D'autres approches ne cherchent pas une classification liée à une explication d'un biais, mais proposent de rattacher le biais à une tâche de haut niveau. L'objectif d'une telle taxonomie est donc différente des taxonomies fondées sur les modèles. Elle ne cherche pas à expliquer un biais mais à identifier la tâche sur laquelle le biais peut apparaître. Ces taxonomies sont donc très utiles pour repérer dans un domaine particulier les tâches où un risque non négligeable d'une prise de décision erronée existe. De ce fait, il existe de très nombreuses taxonomies basées sur les tâches pour de très nombreux domaines comme la visualisation d'information [Dimara et al., 2020], la médecine [Zhang et al., 2004], le management [Carter et al., 2007]... Notre but dans cette section n'est pas d'être exhaustif sur l'ensemble de ces taxonomies. Néanmoins, nous allons nous attarder plus précisément sur la taxonomie de [Arnott, 2001] qui est la plus proche de notre domaine et utilise un vocabulaire proche de l'informatique. En effet, Arnott s'est inspiré des composants d'un système d'aide à la prise de décision pour définir les tâches de haut niveau afin de comparer les violations du raisonnement humain par rapport à une machine. Arnott distingue 6 catégories de biais :

- les *biais de mémoire* qui traitent le rappel d'information et le stockage d'information erronées ;
- les *biais statistiques* qui abordent les traitements d'information erronée par rapport aux lois statistiques ;
- les *biais de confiance* qui augmentent la confiance de l'agent dans sa décision ;
- les *biais d'ajustement* qui traitent de l'ajustement d'une valeur par rapport à une valeur initiale ;
- les *biais de présentation* qui abordent la perception et le traitement d'informations de manière erronée ;
- les *biais de situation* qui traitent de la réaction d'un agent à la situation générale de décision.

37 biais sont ensuite classés dans une de ces catégories. Par exemple, le *biais de confirmation* fait partie des *biais de confiance*, car le biais de confirmation réduit la recherche d'information de la prise de décision en cours à des informations qui sont en accord avec la prise de décision de l'agent. Par conséquent, la confiance de l'agent augmente.

L'ensemble des taxonomies de tâches dans la littérature présente un grand intérêt dans l'étude et la compréhension des biais dans le contexte d'un domaine d'application. En utilisant un vocabulaire se rattachant à un domaine spécifique, comme la prise de décision chez Arnott avec les termes "statistiques", "présentation", "confiance", etc, il est plus facile que les experts de ce domaine se mettent d'accord et comprennent les termes utilisés pour classer les biais. De ce fait, l'application de la taxonomie de biais dans le domaine en question est facilitée. Cet avantage est aussi la limitation de ces taxonomies [Ceschi et al., 2019] : le vocabulaire étant différent, il existe des noms de catégories différents dans les taxonomies qui sont pourtant similaires (*e.g* la catégorie *point de référence* chez Carter et *biais d'ajustement* chez Arnott). La comparaison de ces taxonomies devient alors difficile. Au-delà du problème du vocabulaire utilisé, la comparaison de ces taxonomies est difficile, car elles sont parfois de nature différente. Par exemple, l'erreur de conjonction et l'oubli de la fréquence de base sont dans la catégorie *biais statistiques* chez Arnott mais dans deux catégories différentes chez Carter avec respectivement la catégorie *illusion de contrôle* et *taux de référence*. Enfin, la plupart de ces taxonomies ne prennent pas en compte qu'un biais cognitif peut faire partie de plusieurs catégories en classifiant chaque biais dans une unique catégorie. Ce manque de consensus entre ces taxonomies, cette difficulté de comparaison est là encore un argument supplémentaire à la définition d'une taxonomie formelle des biais.

II.3.2.c Taxonomies empiriques

Les approches empiriques consistent à proposer à un ensemble de participants une batterie de tâches qui sont connues dans la littérature pour être sujettes aux biais cognitifs. Une telle tâche peut-être par exemple *le problème de Linda* [Tversky et al., 1983] qui consiste à répondre à la question suivante :

> Linda a 31 ans, elle est célibataire, franche et très brillante. Elle possède une maîtrise de philosophie. Étudiante, elle se montrait très préoccupée par les questions de discrimination et de justice sociale, elle participait aussi à des manifestations antinucléaires.
>
> Selon vous, Linda a-t-elle plus de chance de :
> (a) être une enseignante de l'école primaire ;
> (b) travailler dans une librairie et prendre des cours de yoga ;
> (c) être activiste dans un mouvement féministe ;
> (d) être une assistante sociale psychiatrique ;
> (e) être membre de la Ligue des femmes votantes ;
> (f) être guichetière dans une banque ;
> (g) être une assureur ;
> (h) être guichetière dans une banque et active dans un mouvement féministe.

Une grande majorité de personne répond à cette tâche avec comme ordonnancement de probabilité : (c) > (h) > (f) ce qui équivaut à *une erreur de conjonction* car (h) est composé des deux évènements indépendants (c) et (f), par conséquent (h) ne peut être qu'inférieur ou égale à (c) ou (f). À partir de ce type de tâche, les approches empiriques évaluent le degré de déviation de la réponse par rapport à la solution normative.

Les approches empiriques font l'hypothèse qu'il existe des différences individuelles inconnues chez les participants (*e.g.* l'état émotionnel actuel du participant, ses compétences cognitives, etc) qui interviennent dans le degré de déviation de la réponse (par rapport à la réponse correcte). Le but est alors de trouver un nombre de facteurs optimaux qui permet d'expliquer la variation dans le degré de réponse pour chaque tâche (et donc chaque biais). Puis de les interpréter à l'aide d'une taxonomie fondée sur un modèle (voir au II.3.2.a).

De nombreux travaux trouvent une solution à deux facteurs qu'ils rattachent à une interprétation différente. Ainsi, [Slugoski et al., 1993] interprète la solution trouvée sous le prisme du modèle d'heuristique de [Tversky et al., 1974] avec les labels *Availability* et *Representativeness*, là où [Weaver et al., 2012] interprète la solution sous le prisme du *dual-process model* de [Kahneman, 2003] avec les labels *Coherence* et *Correspondence*. De ce fait, Slugoski et al. considèrent que la majorité des différences individuelles sont liées à la mémoire (*Availability*) et à la représentativité des croyances (*Representativeness*). Weaver et Stewart considèrent qu'elles sont liées à un jugement imprécis (une erreur dans le processus de *Correspondence*) ou à un jugement incohérent (une erreur dans le processus de *Coherence*).

Enfin, [Ceschi et al., 2019] trouve une solution à trois facteurs qu'il interprète avec les labels tirés des travaux de [Stanovich, 2009, Tversky et al., 1974] :

— *Mindware gaps* qui correspond à la sous-utilisation ou à l'absence de stratégies de raisonnement telles que le raisonnement probabiliste ou l'inférence logique ;
— *Valuation bias* qui correspond à l'évitement d'une perte ou l'attrait à un gain ;
— *Anchoring and Adjustement* qui correspond à l'ajustement d'une valeur par rapport à une valeur initiale.

L'avantage de ces taxonomies est la prise en compte de la multicausalité, c'est-à-dire qu'un biais peut appartenir à plusieurs catégories. Par exemple, le *Framing effect* (*i.e* la tendance à choisir une option en fonction de sa connotation positive ou négative) est à la fois dans *Valuation bias* et *Anchoring and Adjustement* chez Ceshi. En effet, le *Framing effect* peut correspondre à la fois à éviter une connotation négative (une perte) tout en interprétant la valeur négative ou positive de la connotation en fonction d'une référence. De plus, cette approche empirique permet de valider les dimensions importantes dans les taxonomies fondées sur les modèles. Néanmoins, ces approches empiriques n'utilisent pas toutes les mêmes tâches, ce

qui empêche une généralisation des facteurs trouvés. De plus, ces approches sont limitées par le nombre de tâches analysées (généralement peu) ce qui limite le nombre de biais qui peuvent être classés dans la taxonomie résultante. Là encore, le manque de cohérence de ces approches est une limitation à la définition d'un cadre formel à partir de ces approches.

II.3.3 Conclusion

Nous pouvons voir à travers cette littérature des taxonomies des biais cognitifs qu'il existe différentes méthodes pour classer les biais. Cela peut passer par :

- un modèle des heuristiques utilisées par les humains ou un modèle cognitif du raisonnement humain ;
- une classification par les tâches de haut niveau où les biais sont détectés ;
- une classification empirique des biais selon les différences individuelles.

Toutes ces taxonomies sont donc des outils importants pour celui qui cherche à comprendre les erreurs humaines en offrant des caractéristiques d'erreurs à rechercher. Toutefois, au-delà des limitations propres à chacune de ces taxonomies, une limitation importante commune à toutes les approches est à relever :

> Il n'existe **pas de consensus** sur une taxonomie permettant de classifier les biais selon des **caractéristiques précises**.

Les caractéristiques sont imprécises comme le note [Gigerenzer et al., 2009] car elles n'ont pas de délimitations bien définies. Par exemple chez [Arnott, 2001] avec la catégorie *biais de confiance*, qu'est-ce que l'on entend par confiance, qu'est-ce qui augmente la confiance ? ou encore chez [Tversky et al., 1974] avec la notion de degré de disponibilité dans l'heuristique *Availability* qui ne peut pas être vraiment être réfutée. En effet, nous pouvons toujours dire qu'une information est plus disponible en mémoire, car elle est récente (*i.e effet de récence*), quel a eu un impact émotionnel fort (*i.e negativity bias*), si l'information est au début d'une liste (*i.e effet de primauté*), etc. Nous pouvons donc englober autant de notions que l'on veut dans cette notion floue de disponibilité. Par conséquent, il est normal qu'aucun consensus ne puisse être établi sur ces notions dans les différentes approches et que la littérature n'offre pas de caractéristiques précises à retrouver dans le raisonnement d'un agent pour déterminer un biais. De ce fait, nous pensons que la littérature de la taxonomie des biais et plus généralement des sciences cognitives profiteraient grandement d'une taxonomie basée sur des caractéristiques formelles qui permettrait de définir des notions précises pour classer les biais. Nous ne prétendons pas que nos travaux dans cette thèse permettent d'offrir une taxonomie formelle complète des biais au vu de tous les facteurs à prendre en compte (affectif, social, mémoire, etc.) mais c'est un premier pas en ce sens.

II.4 Conclusion

La littérature des erreurs humaines dans le domaine de la sécurité met en avant un point essentiel. Pour expliquer une erreur humaine dans une situation de prise de décision, la clé est de se placer du point de vue de l'humain étudié et rechercher ses croyances à un instant donné. Diagnostiquer une erreur de prise de décision dans une situation d'accident doit donc passer par un modèle modélisant le point de vue de l'agent et capable de déterminer à partir de là les croyances de l'agent. Toutefois, retrouver les causes d'une prise de décision erronée ne suffit pas pour comprendre la prise de décision d'un agent. C'est pourquoi à cela s'ajoute la littérature des sciences cognitives qui met en avant le rôle important des biais cognitifs dans une prise de décision erronée. Ces biais sont le résultat d'heuristiques que l'humain utilise pour compenser sa mémoire et son temps limité. Ces biais permettent de comprendre pourquoi, d'un point de vue cognitif, un agent humain a fait un choix qui peut être basé sur ses émotions, ses croyances passées, un raisonnement simplifié, etc. Un diagnostic d'une prise de décision doit aussi prendre en compte ces facteurs. Cet état de l'art de l'erreur humaine dans les sciences humaines dessine à première vue un modèle en deux temps :

(1) déterminer les causes d'une prise de décision erronée ;

(2) comprendre ces causes par la recherche de biais cognitifs.

Néanmoins, dans le même temps, la littérature sur les biais cognitifs montre qu'il n'existe pas à ce jour un consensus de taxonomie permettant de classifier les biais selon des caractéristiques précises, même au sein des approches de classification similaires. La représentation formelle de ces biais semble une question complexe et est donc une question de recherche pertinente.

III - Modèles informatiques pour la modélisation du raisonnement et l'étude de l'erreur humaine

Nous avons pu voir chapitre II qu'expliquer une erreur humaine dans une situation d'accident consiste selon [Dekker, 2006] à retrouver les croyances rationnelles avec la décision de l'agent humain. Pour cela, il est nécessaire de partir du point de vue de l'agent, de déterminer les décisions incohérentes puis de retrouver les croyances rationnelles. Toutefois, le modèle proposé par Dekker reste vague sur certains concepts comme le "point de vue" de l'agent dont on ne sait pas de quoi il est composé (croyances ? but ? observations ?, etc.) ni comment nous passons concrètement du point de vue et de la décision incohérente aux croyances rationnelles. Il y a donc un manque de formalisation du modèle pour diagnostiquer les erreurs humaines dans une situation d'accident. Or une telle formalisation permettrait d'enlever les ambiguïtés sur les concepts utilisés par le modèle et la méthode de calcul des croyances de l'agent humain.

De plus, nous avons pu voir toujours dans le chapitre II que déterminer les croyances n'était pas suffisant et qu'il était nécessaire de s'appuyer sur les biais cognitifs. Malheureusement, là aussi, la littérature sur les taxonomies des biais cognitifs relève une ambiguïté et des définitions floues qui empêchent de définir des caractéristiques claires pour déterminer les biais cognitifs dans une situation d'accident. De ce fait, une formalisation des biais cognitifs permettrait de répondre à ces limitations.

Nous avons donc un manque de formalisation de la tâche du diagnostic de l'erreur humaine, à la fois sur la méthode, mais aussi sur les processus cognitifs qui permettent de comprendre l'erreur. Or le domaine de l'intelligence artificielle s'intéresse à ce type de problématique en cherchant à formaliser de manière générale le raisonnement humain, notamment, en se reposant sur des langages logiques. C'est pourquoi, pour formaliser le diagnostic de l'erreur humaine, il est nécessaire de se poser la question dans le domaine de l'intelligence artificielle de :

(1) la représentation logique d'un agent qui a des croyances et qui prend des décisions dans un monde dynamique

afin de pouvoir représenter le point de vue de l'agent et ses croyances possibles. Cette problématique très générale englobe plusieurs sous-problématiques que nous allons donc développer dans ce chapitre :

— la représentation du monde dynamique en logique ;
— la représentation des croyances en logique ;
— l'inertie des croyances ;
— les différents changements possibles pour les croyances.

De plus, l'agent que nous cherchons à représenter prend des décisions qui peuvent être erronées. Il est donc nécessaire de se poser la question du :

(2) diagnostic d'un système logique

afin de déterminer les causes de l'erreur de prise de décision. C'est pourquoi dans un deuxième temps, dans ce chapitre, nous allons développer toutes les différentes approches possibles pour cette tâche de diagnostic et mettre en avant un lien entre celle-ci et une des problématiques de (1).

Enfin, nous avons vu section II.2 et section II.3 que les biais cognitifs étaient nécessaires pour trouver le pourquoi du comment du diagnostic, ce qui nous amène à la question de :

(3) la représentation des biais cognitifs dans un cadre formel

afin que le modèle développé dans cette thèse puisse offrir à la fois le comment et le pourquoi de l'erreur dans la prise de décision. Nous allons donc dans ce chapitre développer les problématiques (1), (2) et (3) et voir les différentes solutions proposées dans la littérature.

III.1 Représentation des croyances et des actions

L'une des questions centrales de l'intelligence artificielle depuis ses prémices est de représenter, à l'aide de la logique, un agent qui prend une décision dans un monde dynamique [McCarthy et al., 1969]. Cette question relève une problématique double, car il faut à la fois :

— représenter la dynamique du monde, c'est-à-dire comment le monde évolue par lui-même ou en fonction des actions des agents ;
— représenter la dynamique des croyances des agents, c'est-à-dire comment les croyances de l'agent évoluent en fonction de ce qu'il reçoit comme information sur le monde et de ses croyances.

Nous allons présenter dans cette section les grandes familles de cadre proposées pour représenter la dynamique du monde et la dynamique des croyances dans le temps.

III.1.1 Définitions des concepts de la littérature

La modélisation du raisonnement dans un système dynamique dans la littérature partage des concepts généraux communs. Nous allons dans cette section présenter ces concepts clés qui sont à la base des problématiques et recherches de ce domaine. Pour cela, nous allons nous inspirer largement du travail de synthèse déjà réalisé par [Dupin De Saint-Cyr et al., 2014] :

- Le *modèle* d'un système dynamique est composé à la fois d'un modèle de la dynamique du monde (*e.g* comment le monde change en fonction des actions des agents) et d'un modèle de l'agent (*e.g* quelles sont les croyances de l'agent sur le monde).
- L'*horizon* du modèle est l'ensemble des pas de temps du modèle et peut être infini ou fini.
- Un *état* est une description du système à un pas de temps de l'*horizon* qui contient à la fois la description du monde et l'*état de croyance* de l'agent.
- Un *état de croyance* est une description de la croyance de l'agent sur l'*état* du système.
- Une *trajectoire d'états* est une suite d'*états*, où chaque *état* est indexé par un pas de temps de l'*horizon* et passe à l'*état* suivant par une *transition*.
- Une *transition* décrit formellement les règles de changement du modèle du monde et de l'*état de croyance* en fonction des *évènements*.
- Un *évènement* permet de faire évoluer le système. Plusieurs types d'*évènement* sont possibles :
  - Les *actions* de l'agent qui peuvent modifier à la fois le monde (*actions ontiques*) ou les croyances de l'agent (*actions épistémiques*).
  - Les *observations* de l'agent qui sont un sous-ensemble de l'état du monde auquel l'agent a accès et qui modifient ses croyances.

Ces concepts permettent de donner une idée globale de ce que l'on attend d'un système représentant un agent qui prend des décisions dans un monde dynamique. Toutefois, chacun de ces concepts a ses propres sous-problématiques, par exemple, les actions peuvent être concurrentes et/ou avoir une durée [Sandewall, 1995], les agents peuvent avoir des croyances à propos des croyances des autres agents (*i.e* théorie de l'esprit [Van Ditmarsch et al., 2007]), dans un contexte multi-agent tous les agents n'observent pas les mêmes informations (*i.e public et private announcement* [Baltag et al., 1998]), etc.

Nous allons dans ce chapitre nous intéresser plus précisément aux concepts et problématiques liés à la représentation de l'*état de croyance*, car nous recherchons à formaliser et à construire les croyances de l'agent pour formaliser le diagnostic de l'erreur humaine (voir chapitre II). De plus, nous développerons les problématiques liées au concept de *transition* qui est central pour l'évolution des croyances de l'agent dans le temps et par conséquent un mécanisme qui participe aussi à la construction des croyances de l'agent.

III.1.2 Transition de l'état du monde

La représentation de la dynamique du monde consiste à décrire formellement les effets des actions de l'agent ou d'évènements extérieurs sur un système dynamique afin de pouvoir raisonner dessus. Pour représenter de tels changements, il existe plusieurs cadres de formalisation logique dans la littérature permettant de répondre à cet objectif. Nous développerons dans un premier temps les grandes familles de formalisation logique pour représenter la dynamique du monde. Nous verrons ensuite que ces cadres formels doivent répondre à trois problématiques principales (*décor, qualification et ramification*) et présenterons certaines solutions proposées.

III.1.2.a Yale Shooting Problem

Avant de définir les différents cadres de formalisation, nous proposons de les illustrer avec un exemple connu de la littérature, le *Yale Shooting Problem* (YSP) de [Hanks et al., 1987] où un agent décide de charger son pistolet et tirer sur autre personne. Afin que cette thèse soit adaptée à tout public, nous remplacerons la personne par une dinde.

YSP se déroule sur trois pas de temps (voir figure III.1) :

— Dans la situation initiale, la dinde est vivante (*alive*) et le pistolet est déchargé (*not loaded*). L'agent décide de faire l'action de charger son pistolet (*Load*) qui a pour conséquence que le pistolet soit chargé au pas de temps suivant (voir figure III.1a).

— Au pas de temps suivant, l'agent décide d'effectuer une action d'attente ($Wait$) qui n'a aucune conséquence sur le monde (voir figure III.1b).

— Au dernier pas de temps, l'agent décide de tirer $Shoot$ ce qui a pour conséquence de tuer la dinde avec un pistolet chargé (voir figure III.1c).

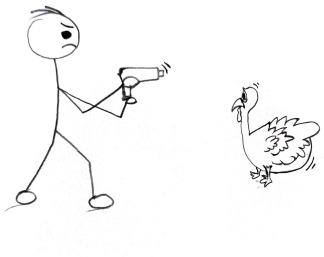
Nous allons reprendre cet exemple tout le long de cette section et verrons les différentes formalisations possibles de ce problème. Il est possible que nous utilisions des petites variations de ce problème afin qu'il s'adapte à une problématique donnée. Si tel est le cas, nous le préciserons dans la suite.

III.1.2.b Cadres de formalisation

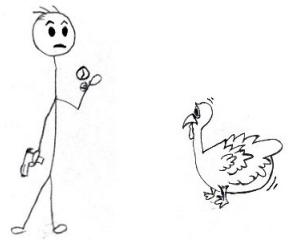
Le calcul des situations Le premier cadre historique qui a été proposé est celui du *calcul des situations* par [McCarthy et al., 1969] qui a été ensuite amélioré par [Reiter, 1991]. C'est la version de ce dernier que nous présentons ici en l'illustrant avec YSP (voir au III.1.2.a). Le *calcul des situations* est un langage logique de premier ordre composé :

— D'un ensemble d'actions où chaque action est un prédicat logique. Par exemple, le prédicat $Load(g)$ correspond à l'action de charger le pistolet g.

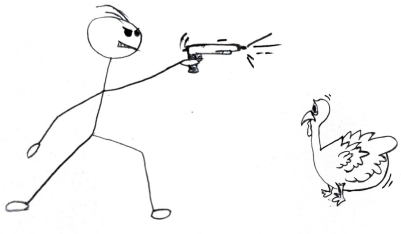
— D'un ensemble de prédicats logiques pouvant changer dans le temps appelés *fluents*. Par exemple le prédicat $loaded(g, s)$ qui est vrai si le pistolet g est



(a) Situation initiale



(b) Deuxième pas de temps



(c) Troisième pas de temps

Figure III.1 – Yale Shooting Problem

chargé dans la situation s.

— D'un ensemble de situations qui représente un historique d'actions effectuées. Cet historique est représenté par une séquence de fonction $do(a, s)$ qui retourne la situation s' qui résulte de l'action a dans la situation s. Par exemple, une situation possible peut-être :

$$do(Shoot(g), (do(Wait, (do(Load(g), S_0))))$$

où l'agent depuis la situation initiale S_0 charge le pistolet g, attend, puis tire avec.

À partir de ces éléments, la dynamique du monde est décrite par des formules logiques décrivant les préconditions et effets des actions. Les préconditions d'action sont décrites grâce au prédicat $Poss(a, s)$ qui doit être vrai si l'action a peut être effectué dans la situation s. Par exemple :

$$Poss(Shoot(g), s) \leftrightarrow loaded(g, s)$$

c'est-à-dire que pour tirer avec le pistolet g, il faut que celui-ci soit chargé. De la même manière, les effets d'une action sont décrites par des formules logiques si la précondition de l'action est respectée dans la situation. Par exemple :

$$Poss(Shoot(g), s) \rightarrow \neg loaded(g, do(Shoot(g, s)))$$

si l'agent peut tirer avec g dans la situation s et que celui-ci l'effectue alors le pistolet g n'est plus chargé.

Le *calcul des situations* offre un cadre très expressif pour représenter la dynamique des actions du fait de l'utilisation d'une logique de prédicat et par conséquent la possibilité d'utiliser des quantificateurs (*i.e* $\exists, \forall$) dans la logique. Toutefois, son expressivité implique une plus grande complexité d'implémentation. En effet, l'implémentation du raisonnement sur les formules logiques passe généralement par des solveurs SAT ou ASP (Answer Set Programming) qui utilisent une logique propositionnelle. Or le *calcul des situations* utilise aussi des fonctions (comme la fonction do) et des quantificateurs qui nécessitent alors un travail de translation entre le cadre du *calcul des situations* et la logique propositionnelle [Lee et al., 2010]. Pour éviter cette complexité, d'autres recherches se sont intéressées à la représentation de la dynamique du monde dans un cadre purement propositionnel.

Langage d'actions Les langages d'actions sont très utilisés dans la communauté de la planification. Leur principe est de représenter les actions comme des transitions entre deux états (voir figure III.2). La transition que représente l'action est exprimée par des règles d'effets et de précondition de l'action qui définissent les propositions qui vont changer dans l'état S_{t+1} et quelles conditions sont nécessaires dans l'état S_t pour que l'action a soit effectuée.

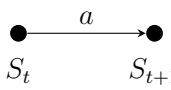


Figure III.2 – Transition de l'état S_t à S_{t+1} par l'action a

Le langage d'actions le plus élémentaire est STRIPS [Fikes et al., 1971] et utilise un opérateur logique de contrainte, que nous notons $\Rightarrow$, qui lie l'état du monde avant et après une action a pour représenter les changements. L'opérateur pour une action a est de la forme $c \Rightarrow e$ où e devient vrai après l'action a si et seulement si c est vrai. Par exemple, si nous voulons représenter l'action de tirer dans YSP, nous avons :

$$loaded \Rightarrow \neg alive$$

C'est-à-dire que l'action de tirer, rend la proposition $alive$ fausse si $loaded$ est vrai. Il est à noter que l'opérateur $\Rightarrow$ est différent de l'opérateur d'implication logique classique noté $\rightarrow$, c'est-à-dire que la contraposition logique n'est pas vraie : $loaded \Rightarrow \neg alive$ n'est pas équivalent à $alive \Rightarrow \neg loaded$. En n'utilisant pas l'implication classique, le langage STRIPS perd de l'expressivité du fait qu'il ne peut pas exprimer des règles d'inférence et doit écrire tous les effets possibles d'une action. Par exemple, considérons une variation de YSP où nous introduisons le fait que tirer avec le pistolet fait du bruit et par conséquent effraie les oiseaux. Bien que ce soit le bruit qui cause la peur des oiseaux, nous sommes obligés dans le langage STRIPS de spécifier que l'action de tirer va provoquer ces deux effets :

$$loaded \Rightarrow \neg alive \wedge noise \wedge fear$$

Nous verrons que ce problème est connu dans la littérature comme *le problème de la ramification* dans la prochaine section. Pour palier ce problème d'expressivité, d'autres langages d'actions ont été développés dans la littérature.

La plupart des autres approches des langages d'actions se basent sur une *implication causale* [Gelfond et al., 1993, Geffner, 1990, Turner, 1999] qui a pour but de distinguer les formules logiques qui sont vraies, des formules qui sont vraies pour une raison (une cause). Syntaxiquement, une implication causale :

$$\varphi \leftarrow \psi$$

se lit "il y a une cause pour φ si ψ est vraie", avec φ et ψ des formules de la logique classique. De ce fait, les effets et préconditions d'actions peuvent être exprimés sous forme d'implication causale sous la forme *effets* $\leftarrow$ *actions et préconditions*. Par exemple :

$$\neg alive \leftarrow (loaded \wedge Shoot)$$

indique que le fait que la dinde ne soit plus en vie est causé par le fait de tirer en ayant un pistolet chargé. En utilisant un tel opérateur d'implication, il est possible d'exprimer les inférences sur les effets d'une action. Par exemple :

$$(\neg alive \wedge noise) \leftarrow (loaded \wedge Shoot)$$
$$fear \leftarrow noise$$

Le bruit est causé par le fait de tirer avec le pistolet et la peur des oiseaux est causé par le bruit.

Ce type de langage a l'avantage d'être facilement transposable sur un solveur du fait de l'utilisation de la logique propositionnelle, et ils ont donc une implémentation plus facile au prix de l'expressivité du calcul des situations (voir le calcul des situations).

Logique dynamique La logique dynamique qui était à l'origine utilisée pour raisonner sur l'exécution de programmes informatiques [Harel et al., 2001] montre une utilisation possible dans le cadre du raisonnement sur les actions des agents. Ces logiques dynamiques reposent sur les opérateurs booléens classiques et des opérateurs modaux de la forme $[a]\varphi$. Cet opérateur se lit "φ est vrai après l'action a". Par exemple :

$$loaded \rightarrow [Shoot]\neg alive$$

indique que si le pistolet est chargé, cela implique que si l'action de tirer est effectuée, alors la dinde est morte. L'opérateur de la forme $< a > \varphi$ est une abréviation de $\neg[a]\neg\varphi$, c'est-à-dire qu'il est possible que p soit vrai après l'action a.

En plus de prendre en compte le changement, la logique dynamique permet d'exprimer des changements plus complexes avec différentes opérations sur les actions :

— *les séquences* notées $a;b$ qui expriment le fait que l'action b est effectuée à la suite de a. Ainsi $[a;b]\varphi$ indique que φ est vrai après avoir effectué l'action a puis b.
— *les choix* notés $a \cup b$ qui expriment le fait que l'action a ou b peut être effectuée. Ainsi $[a \cup b]\varphi$ indique que φ est vrai après avoir effectué l'action a ou b
— *les itérations* notées $a*$ qui expriment le fait que l'action a est effectué 0 fois ou plusieurs fois de manière séquentielle. Ainsi $[a*]\varphi$ indique qu'après avoir effectué l'action a de manière répétée, φ est vrai.

Nous verrons dans la suite de ce chapitre qu'une partie de la littérature s'est concentrée sur l'extension de cette logique dans un cadre épistémique avec la *Dynamic Epistemic Logic* [Baltag et al., 1998].

La logique dynamique, en se basant sur une logique modale, permet d'exprimer des changements complexes en considérant des choix, des séquences et des itérations d'actions. Cette expressivité va bien sûr au détriment de la complexité de la logique, qui est plus complexe qu'une logique propositionnelle d'actions [Aucher et al., 2013].

III.1.2.c Conclusion sur les cadres de formalisation

La littérature sur la formalisation de la dynamique des actions sur le monde met en avant plusieurs solutions possibles. Chacune de ces solutions a ses avantages et inconvénients, et le choix de préférer une solution par rapport à une autre repose sur le but recherché. Si nous voulons exprimer des changements complexes avec des actions concurrentes, des choix, etc, il est préférable de se tourner vers une logique dynamique. Si nous voulons exprimer, de manière simple, une situation complexe, il est préférable de se tourner vers le calcul des situations qui utilise les prédicats logiques. Si nous souhaitons favoriser l'implémentation à l'expressivité du langage, il est préférable de se tourner vers une logique d'actions.

Ces différentes logiques, au-delà de leurs différences, doivent répondre à des difficultés inhérentes aux problèmes de changements du monde par les actions. Nous allons dans la prochaine section passer en revues ces difficultés et les solutions proposées.

III.1.3 Trois difficultés pour la dynamique du monde

Au-delà du cadre utilisé pour représenter la dynamique du monde, la littérature met en avant que représenter les changements des actions dans un système logique pose trois grandes difficultés.

III.1.3.a Problème du décor

Le *problème du décor* (*frame problem* en anglais) a été introduit en 1969 par [McCarthy et al., 1969]. C'est un problème très connu en modélisation logique qui peut être décrit comme la difficulté à représenter les effets d'une action en logique sans avoir à représenter explicitement tous les non-effets évidents [Shanahan, 2016]. Pour illustrer ce propos, considérons une formalisation simple, en logique du premier ordre, du *Yale Shooting Problem* (YSP) introduit au III.1.2.a :

$$\begin{array}{lcl} \textit{init} & = & \mathrm{alive}(0), \neg\,\mathrm{loaded}(0) \\ \textit{actions} & = & \mathrm{Load}(0), \mathrm{Wait}(1), \mathrm{Shoot}(2) \\ \textit{rules} & = & \mathrm{Load}(0) \rightarrow \mathrm{loaded}(1), \\ & & (\mathrm{Shoot}(2) \wedge \mathrm{loaded}(2)) \rightarrow \neg\,\mathrm{alive}(3) \end{array}$$

Les prédicats dans *init* décrivent la situation initiale : la dinde est vivante et le pistolet est chargé. Les prédicats dans *actions* modélisent le déroulement des actions (l'agent charge le pistolet, attend et tire) et les prédicats dans *rules* modélisent la physique du monde : tirer sur la dinde avec un pistolet chargé, la tue.

Avec cette modélisation, le bon sens nous dit que $\neg\,\mathrm{alive}(3)$ est vrai. Toutefois, les effets de l'action *Wait* ne sont pas définis dans le modèle et ne peuvent être déduits sans des règles logiques qui indiquent que $\mathrm{loaded}(t+1) = \mathrm{loaded}(t)$ quand rien d'autre ne dit le contraire, c'est-à-dire une *inertie* sur les propositions du monde. Écrire ces règles manuellement pour chaque fait initial et toutes les actions possibles est une charge trop lourde dans le cas général. C'est ce qu'on appelle le *problème du décor*. Plusieurs solutions au *problème du décor* existent. Nous n'allons aborder ici que les solutions principales de la littérature.

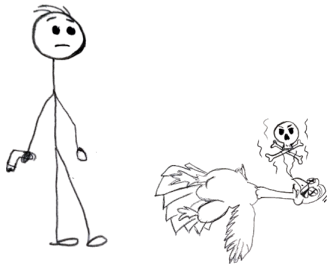
La *circumscription* de [McCarthy, 1986] est la première solution proposée et consiste à sélectionner les modèles (c'est-à-dire les ensembles de propositions) qui minimisent le nombre de changements dans le monde. L'idée est que par défaut les propositions logiques conservent leur état, sauf changement forcé par l'action. Toutefois, cela ne marche pas sur le YSP où deux solutions minimales peuvent être trouvées : la solution conforme à notre intuition, dans laquelle $\mathrm{loaded}(2) \wedge \neg\,\mathrm{alive}(3)$ est vraie, mais aussi une solution contre-intuitive $\neg\,\mathrm{loaded}(2) \wedge \mathrm{alive}(3)$ dans laquelle le pistolet se décharge magiquement pendant l'action *Wait* (voir figure III.3).

La solution de [Reiter, 1991] (*succesor state axiom*) permet de résoudre ce problème en ajoutant des axiomes qui pour chaque proposition décrivent comment une proposition devient vraie. Chaque axiome décrit deux possibilités :

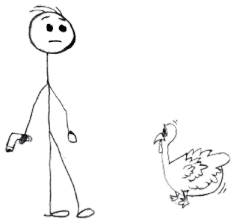
(1) si la proposition était fausse, quelles actions la rendent vraie ;
(2) si la proposition était vraie, quelles actions ne la rendent pas fausse.

Si nous reprenons l'exemple ci-dessus, la proposition $\mathrm{loaded}(1)$ est vraie si :

(1) $\mathrm{loaded}(0)$ est fausse et que $\mathrm{Load}(0)$ est effectuée ;
(2) $\mathrm{loaded}(0)$ est vraie et que toute autre action que $\mathrm{Shoot}(0)$ est effectuée.



$\text{loaded}(2) \wedge \neg\,\text{alive}(3)$



$\neg\,\text{loaded}(2) \wedge \text{alive}(3)$

Figure III.3 – Deux solutions minimales

Enfin [Sandewall, 1998] propose une solution à base de prédicat de type "occlusion" de la forme $Occlude(t, p)$ qui exprime qu'au temps t la proposition p est exemptée de l'inertie par défaut des propositions. Par exemple :

$$(\text{Load}(0) \wedge \neg\,\text{loaded}(0)) \rightarrow Occlude(1, \text{loaded})$$

exprime que faire l'action de charger un pistolet non chargé casse l'inertie de la proposition loaded.

La solution de Reiter a été largement adoptée dans les différentes formalisations des changements du monde. C'est dans ce cadre de formalisation que Reiter a développé sa solution au *problème du décor* à base d'axiomes. Si nous voulons exprimer les changements possibles pour le *fluent* $loaded(g, s)$ dans *le calcul des situations*, nous avons pour tout g, s, a $loaded(g, do(a, s))$ qui équivaut à :

$$((\neg loaded(g, s) \wedge a = Load(o)) \vee (loaded(g, s) \wedge a \neq Shoot(g))$$

Dans le cas des logiques d'actions basées sur l'*implication causale*, les règles causales sont traduites sous forme de transition où chaque proposition est indicée temporellement avec φ_t qui représente φ à l'instant t avant l'action effectuée au pas de temps t et φ_{t+1} représente φ après que l'action du pas de temps t a été effectuée [Gelfond et al., 1993]. Ainsi, φ est vrai à $t + 1$ si :

(1) φ est vrai à t et il n'existe pas de règle causale vraie dont la conclusion est $\neg\varphi_{t+1}$;

(2) $\neg\varphi$ est vrai à t et il existe une règle causale vraie dont la conclusion est φ_{t+1}.

Ce qui revient à la solution de Reiter pour le *problème du décor*. Si nous reprenons YSP, nous avons :

$$\begin{array}{ll} alive_{t+1} & \equiv \quad alive_t \wedge \neg(loaded_t \wedge Shoot_t) \\ \forall t < n & \end{array}$$

où n est le nombre de pas de temps de la situation.

Enfin, dans le cas de la logique dynamique, [Van Ditmarsch et al., 2011] montrent là aussi que la solution de Reiter est possible.

Le *problème du décor* est donc un problème central dans la représentation des changements d'actions en spécifiant que les propositions du monde ont une *inertie* par défaut qui peut être "cassée" par les actions de l'agent sur le monde. Il existe plusieurs solutions pour palier ce problème dont la plus populaire est la solution de Reiter qui permet de spécifier les axiomes qui rendent une proposition vraie. Ainsi, il est attendu que dans toutes les formalisations logiques cherchant à représenter les changements d'actions, une des solutions du *problème du décor* soit prise en compte.

III.1.3.b Problème de la qualification

Le *problème de la qualification* [McCarthy, 1981] est la difficulté à représenter l'ensemble des préconditions d'une action. Par exemple, dans le YSP, l'action de tirer entraînera la mort de la dinde si et seulement si le pistolet est chargé, mais aussi que la balle chargée n'est pas une balle à blanc, que le pistolet n'est pas en panne, etc. Il y a donc un nombre indéterminable de raisons pour lesquelles tirer avec le pistolet ne tue pas la dinde. Trouver un moyen de ne pas spécifier l'ensemble des contextes pour lesquels une action n'a pas l'effet attendu revient à trouver une solution pour le *problème de la qualification*.

Le *problème de la qualification* est principalement un problème de modélisation. En effet, le processus de modélisation consiste à abstraire le monde physique et par conséquent en faire une simplification qui se concentre sur un aspect spécifique de la situation étudiée. Le but est donc de retrouver les préconditions d'action qui sont importantes dans le contexte de la situation étudiée. Dans notre cas, celui de l'analyse d'une situation d'accident, l'enquêteur sait déjà quelles actions étaient incohérentes avec les préconditions d'action qui n'ont pas été respectées. Si nous reprenons l'exemple du Rio-Paris, la précondition de ne pas être en décrochage pour effectuer l'action de tirer le manche de contrôle n'a pas été respectée. Le modélisateur va donc simplifier la représentation du monde autour de ces préconditions non respectées par l'agent pour effectuer une action, car il n'est pas nécessaire d'explorer toutes les préconditions non respectées par l'agent qui n'ont pas de sens dans la situation étudiée.

III.1.3.c Problème de la ramification

Le *problème de la ramification* [Finger, 1987] est dual au *problème de la qualification*. C'est la difficulté à représenter l'ensemble des postconditions d'une action. Par exemple, dans YSP, le fait de tirer au pistolet a pour effet de provoquer une détonation qui aura pour effet de faire fuir les oiseaux aux alentours. Ce deuxième effet n'est pas relié directement à l'action de tirer, mais à une règle qui indique que s'il y a du bruit, alors les oiseaux fuient. Le *problème de la ramification* consiste

alors à trouver une solution pour ne pas décrire explicitement les effets indirects d'une action. De ce fait, tout langage logique permettant de ne pas spécifier toutes les propositions qui doivent changer suite à une action, mais permet de les trouver par inférence logique ne tombe pas dans le *problème de ramification*. Au contraire, le langage d'action STRIPS tombe dans ce problème du fait qu'il ne contient que des contraintes entre deux états et non une implication logique classique. De ce fait, le langage en lui-même ne permet pas de déduire tous les effets indirects d'une action, nous sommes obligés de les spécifier.

Dans notre cas, nous utiliserons un langage logique qui prend en compte l'inférence afin de simuler un raisonnement de l'agent. Enfin, tout comme le *problème de qualification*, le modélisateur connaissant la situation d'accident doit pouvoir se limiter aux croyances de l'agent qui ont une importance sur la situation étudiée et par conséquent n'avoir que les effets indirects importants d'une action.

III.1.3.d Conclusion

Représenter formellement les changements dans le monde suite à des actions n'est pas juste une problématique d'expressivité d'un langage logique, mais aussi la capacité à répondre à des difficultés sur la modélisation de ces changements. L'un des problèmes majeurs étant le *problème du décor* afin de prendre en compte l'*inertie* des propositions du monde et des changements des actions sur cette *inertie*. La littérature présente plusieurs solutions à ce problème dont la solution de Reiter qui s'adapte à chaque logique développée au III.1.2.b. Les problèmes de *la ramification* et de *la qualification* sont quant à eux des problèmes plus liés au choix de modélisation du modélisateur. Dans notre cas, celui de l'analyse de situations d'accident, nous pensons qu'au vu des informations disponibles pour le modélisateur, celui-ci peut se limiter aux effets et préconditions importantes pour l'accident et ainsi éviter ces deux problèmes.

Les changements suite à des actions sont un aspect important d'une situation où un agent prend des décisions. Il faut toutefois prendre en compte d'autres changements qui ne sont pas seulement sur le monde, mais sur les croyances de l'agent. C'est pourquoi nous allons nous intéresser dans la prochaine section à la représentation des croyances de l'agent et de la dynamique de ces croyances.

III.1.4 Les logiques de croyance

Nous avons vu dans la section précédente les formalisations pour la dynamique du monde, c'est-à-dire comment le monde évolue en fonction des actions de l'agent. Nous devons aussi représenter le raisonnement de l'agent afin de notamment déterminer ce que l'agent croit ou ne croit pas dans une situation donnée. La littérature s'est intéressée à ces logiques avec les logiques *doxastiques* (*i.e* le raisonnement sur les croyances) et les logiques *épistémiques* (*i.e* le raisonnement sur les connaissances) initiées par [Hintikka, 1962]. La littérature distingue les croyances

des connaissances par le fait que si l'agent connaît φ alors cela veut dire que φ est vrai dans le monde et que si l'agent croit φ alors φ peut être vrai ou faux dans le monde. Dans cette thèse, nous nous intéressons essentiellement aux croyances de l'agent afin de comprendre son raisonnement. C'est pourquoi nous allons nous concentrer sur l'aspect *doxastique* des logiques sachant qu'une logique peut être à la fois *doxastique* et *épistémique*. Nous allons dans un premier temps définir les concepts généraux de ces logiques puis nous attarder sur les logiques BDI très utilisées en intelligence artificielle.

III.1.4.a Concepts des logiques *doxastiques*

Une logique *doxastique* repose généralement sur :

- Un ensemble de propositions logiques que nous notons $\mathcal{P}$ où chaque proposition a une interprétation spécifique. Par exemple $\varphi \in \mathcal{P}$ peut représenter la proposition "le pistolet est chargé".
- Un ensemble d'agents noté $\mathcal{A}$ qui contient l'ensemble des agents présents dans la situation modélisée.
- Un ensemble d'évènements noté E qui sont des propositions qui changent le monde ou les croyances (*e.g* l'action d'un agent).
- Des opérateurs modaux "épistémiques" permettant de représenter une propriété sur les propositions logiques dans $\mathcal{P}$ en relation avec le raisonnement de l'agent (croyances, désirs, etc). Par exemple, l'opérateur $B_\alpha\varphi$ avec $\alpha \in \mathcal{A}$ peut indiquer que l'agent α croit que φ.
- Des opérateurs modaux propres à la logique utilisée. Par exemple des opérateurs de la logique temporelle ou des opérateurs de changement de modèle.
- Les opérateurs de logiques classiques ($\wedge, \vee, \neg, \rightarrow, ...$) et les propositions $\bot$ et $\top$ qui viennent respectivement définir la proposition constamment fausse et la proposition constamment vraie.

Ces différents ensembles servent de base à une syntaxe et sémantique logique qui diffère en fonction de la logique utilisée. Toutefois, il est attendu que les opérateurs modaux respectent un fonctionnement caractérisé par des axiomes. Par exemple, l'opérateur doxastique B_α de croyance respecte généralement les axiomes KD45 :

(K)	$B_\alpha(\varphi \rightarrow \psi) \rightarrow (B_\alpha\varphi \rightarrow B_\alpha\psi)$	: Fermeture
(D)	$B_\alpha\varphi \rightarrow \neg B_\alpha\neg\varphi$	: Cohérence
(4)	$B_\alpha\varphi \rightarrow B_\alpha B_\alpha\varphi$	: Introspection positive
(5)	$\neg B_\alpha\varphi \rightarrow B_\alpha\neg B_\alpha\varphi$	: Introspection négative

L'axiome (K) indique que le raisonnement de l'agent est fermé par implication, c'est-à-dire que si l'agent croit à l'hypothèse alors il croit à toutes les conclusions de cette hypothèse. Si l'agent croit que s'il tire avec le pistolet, cela a pour conséquence de tuer la dinde, alors s'il croit qu'il tire, il doit croire que la dinde est morte. L'axiome (D) indique que l'agent ne peut pas croire tout et son contraire

(*i.e* croire φ et $\neg\varphi$ en même temps). Enfin, les axiomes (4) et (5) indiquent que l'agent est conscient de ses propres croyances et non croyances.

Suivant ces concepts, plusieurs logiques *doxastiques* ont été définies dont deux très populaires en intelligence artificielle : la *Dynamic Epistemic Logic* (DEL) et la logique *Beliefs Desires Intention* (BDI).

III.1.4.b Dynamic Epistemic Logic

La *Dynamic Epistemic Logic* (DEL) s'intéresse à l'évolution des croyances de l'agent face à l'occurrence d'évènements dans le monde, et est une extension épistémique de la logique dynamique abordée au III.1.2.b. Le développement de cette logique a débuté par les travaux de [Plaza, 1989], suivis de nombreux autres avec par exemple [Baltag et al., 1998, Gerbrandy et al., 1997]. Elle repose sur les principes décrits à la section précédente en ayant un opérateur épistémique de croyance B respectant KD45 et un opérateur modal à la logique dynamique, l'opérateur $[\mathcal{E}, E_0]$. $\mathcal{E}$ correspond à un modèle d'évènement qui définit la dynamique du système, c'est-à-dire l'ensemble des évènements (*e.g* une action, une observation, etc.) et comment ces évènements changent le monde. E_0 est un ensemble d'évènement pointé par le modèle. Nous notons $\mathcal{L}$ le langage logique utilisé par la logique DEL. Un modèle d'évènement $\mathcal{E} = (E, R_i, post, pre)$ est alors composé :

- d'un ensemble non vide d'évènements E ;
- d'une relation d'accessibilité binaire $R_\alpha \subseteq E \times E$, avec $\alpha \in \mathcal{A}$ qui permet d'indiquer les évènements qui sont perceptibles pour l'agent α ;
- d'une fonction $post : E \to (\mathcal{P} \to \mathcal{L})$ qui permet d'indiquer les propositions qui sont mises à jour après un événement (*e.g* la valeur de vérité de φ est assignée à la valeur de $post(e, \varphi)$) ;
- d'une fonction $pre : E \to \mathcal{L}$ qui permet d'indiquer si l'évènement est exécutable ou non.

Ainsi un changement dans les croyances d'un agent suite à un évènement peut se décrire par la formule $[\mathcal{E}, E_0]B_\alpha\varphi$ se lit "L'ensemble des évènements E_0 peut se produire et l'agent α croit que φ est vrai par la suite". Les changements de l'évènement E_0 sur le monde sont pris en compte directement dans le modèle $\mathcal{E}$ avec la fonction *post*.

Les évènements peuvent être de plusieurs types que ce soit des *public announcement* [Plaza, 1989] où tous les agents peuvent observer la vérité de φ, des *private announcement* où seul un sous ensemble d'agents peut observer la vérité de φ ou encore des *semi-private announcement* où un sous ensemble d'agents n'observe pas la vérité de φ mais observe que d'autres agents ont l'information sur la vérité de φ [Belardinelli et al., 2017].

Cette approche a l'avantage de proposer une formalisation des changements à la fois des croyances des agents et du monde dans le temps. Les croyances des agents évoluent en fonction d'évènements qui peuvent être partiellement, pas du tout ou complétement perceptibles. La DEL est donc une logique très adaptée pour

un contexte multi-agents où les agents ont accès à des informations différentes. Par exemple, elle est utilisée pour la formalisation de la *théorie de l'esprit* qui cherche à représenter les croyances d'un agent α sur les croyances des autres agents [Dissing et al., 2020].

Toutefois, la littérature sur la DEL, en se concentrant essentiellement sur l'évolution des croyances, n'aborde pas la construction de plan d'action par les agents. C'est-à-dire quelles actions l'agent planifie et comment il les choisit. Une autre partie de la littérature des logiques des croyances s'est concentrée sur cette question autour du concept d'*intention*. Nous abordons ces logiques dans la prochaine section.

III.1.4.c Les logiques BDI

Les logiques BDI sont des logiques *doxastiques* dont l'objectif principal est de représenter formellement comment un plan d'action se construit en fonction des croyances et des désirs. Pour cela, elles reposent sur une formalisation du modèle BDI.

Le modèle BDI (*Beliefs, Desires, Intentions*) de [Bratman, 1987] permet de représenter le raisonnement d'un agent intelligent dans un environnement dynamique. Le modèle BDI repose sur trois concepts :

- les croyances comme une représentation du monde de l'agent ;
- les désirs comme un état du monde qui est souhaitable pour l'agent et peuvent être incohérents entre eux ;
- les intentions comme des plans d'actions permettant de satisfaire un sous ensemble des désirs.

Les logiques BDI sont très populaires en intelligence artificielle. Cette popularité repose en partie sur les concepts au cœur du modèle BDI qui correspondent facilement au langage que les gens utilisent pour décrire leurs raisonnements et leurs actions [Norling, 2004]. Quatre logiques BDI ont une influence majeure dans la littérature [Meyer et al., 2015, Herzig et al., 2017]. Les trois premières se basent sur la logique KD45 mais différent essentiellement dans la formalisation de l'intention. La quatrième quant à elle propose un cadre de formalisation plus simple avec une approche "à la base de données".

Logique de Cohen et Levesque [Cohen et al., 1990] définissent l'intention par trois notions primitives : les *buts persistants*, les *objectifs à atteindre* et les *choix*. Ainsi l'intention d'une proposition φ pour un agent α est définie comme le fait que l'agent α a pour *but persistant* φ et croit qu'il peut effectuer une action pour satisfaire φ. Un *but persistant* φ pour un agent α est défini comme le fait que l'agent α a l'*objectif à atteindre* φ et gardera cet objectif tant qu'il ne croit pas qu'il est atteint. Un *objectif à atteindre* φ pour un agent α consiste à ce que l'agent α choisit que φ soit vrai alors qu'il croit que φ est pour l'instant faux. Enfin, le *choix* envers φ consiste en un désir qui est considéré par l'agent

comme réalisable. Cette dernière notion de choix est définie comme un opérateur modal dans la logique. Ainsi, les notions d'*objectif*, de *but* et d'*intention* sont définies à partir de l'opérateur de choix, de croyances et des opérateurs temporels de la logique temporelle linéaire (LTL). Cependant, [Herzig et al., 2017] dans leur discussion des limites des logiques BDI, mettent en avant la complexité bien trop grande de la logique de Cohen et Levesque pour être implémentée.

Logique de Rao et Georgeff [Rao et al., 1991] définit les croyances, les désirs et les intentions directement comme des notions primitives. C'est-à-dire que ces trois notions sont représentées par un opérateur modal (*i.e* respectivement $B_\alpha\varphi$, $G_\alpha\varphi$, $I_\alpha\varphi$). De plus, cette logique utilise les opérateurs temporels de la *Computation Tree Logic* (CTL) qui est une logique plus forte que LTL en permettant de raisonner sur des ramifications temporelles. Ces ramifications permettent notamment de résoudre le problème de *Little Nell* [McDermott, 1982] qui consiste à ne pas abandonner l'intention de sauver Nell d'un train qui s'apprête à l'écraser si l'agent croit que Nell sera en sécurité dans le futur. En effet, dans une logique linéaire, il n'existe qu'un seul futur possible avec l'intention de sauver Nell : elle sera en sécurité. De ce fait, il n'y a aucune raison de garder l'intention de sauver Nell car l'objectif va être atteint par ses actions. Or dans le cas d'une logique CTL, il existera toujours une ramification (un futur possible) où Nell n'est pas en sécurité. Ainsi, l'intention de sauver Nell est gardée dans la logique de Rao et Georgeff. Toutefois [Herzig et al., 2017] pose la même critique sur la logique de Rao et Georgeff que celle de Cohen et Levesque : sa trop grande complexité computationnelle.

KARO framework [Van Linder et al., 1998] définissent le framework KARO à l'aide d'une logique dynamique (voir au III.1.2.b). En plus de l'opérateur modal de croyance respectant KD45, il existe un opérateur modal de désir D et un opérateur modal d'aptitude A pour une action. KARO est donc une logique très proche de la *Dynamic Epistemic Logic* mais permet de prendre en compte l'intention, c'est pourquoi nous la classons en tant que logique BDI. À la manière de Cohen et Levesque, à partir des opérateurs modaux de croyance, désir et aptitude, les auteurs proposent une succession de définitions (*i.e.* but, réalisabilité, possibilité et possibilité concrète) permettant de définir finalement l'intention. Cependant, en utilisant une logique "à la DEL", cette logique permet d'avoir des ramifications et non une structure linéaire. Ainsi le framework KARO permet d'avoir les avantages de la logique de Rao et Georgeff tout en ayant une complexité beaucoup moins importante que cette dernière [Aucher et al., 2013].

Approche base de données [Shoham, 2009] diffère des précédentes logiques en proposant une vision proche des bases de données pour la formalisation de l'agent. Dans ce cadre formel, un agent est représenté par deux "bases de données" (B, I) :

— Une base de croyances B qui est un ensemble de formules logiques clôt logiquement où chaque proposition est indicée temporellement. De ce fait,

si $\varphi_t \in B$ alors cela veut dire que l'agent croit φ au temps t.

— une base d'intentions I qui est un ensemble d'actions indicées temporellement. De ce fait, si $a_t \in I$ cela veut dire que l'agent à l'intention d'effectuer l'action a au temps t.

Le modèle de l'agent doit alors respecter quatre contraintes :

(C1) B est cohérent.

(C2) I ne peut contenir deux actions différentes avec le même indice temporel (*i.e.* une seule action à la fois peut être effectuée).

(C3) Si l'agent à l'intention d'effectuer une action, alors il doit croire aux postconditions de cette action.

(C4) Si l'agent a l'intention d'effectuer une action, alors il ne peut pas croire que les préconditions de l'action soient fausses.

Une telle représentation permet d'être moins complexe que les logiques précédentes et permet une implémentation facilitée de l'architecture BDI. Cette approche a été notamment utilisée dans l'application *Timeful* rachetée par Google en 2015 [Shoham, 2015].

Les facteurs humains dans la BDI Les logiques BDI introduisent des notions qui permettent de développer l'expressivité du modèle en termes de cognition en prenant en compte les désirs et les intentions. Sur la base de ces logiques, plusieurs travaux introduisent les émotions qui jouent un rôle important dans la prise de décision. Par exemple, [Adam et al., 2009] modélisent de manière qualitative 20 émotions tirées du modèle OCC [Ortony et al., 1990]. [Dastani et al., 2012] quant à eux, modélisent quatre émotions de manière quantitative en considérant la graduation des croyances, désirs et intentions pour prendre en compte l'intensité des émotions. Nous pouvons donc voir que les logiques BDI sont une base solide pour représenter et comprendre le raisonnement humain.

Toutefois, plusieurs limites sont à considérer tant sur le plan technique que sur le plan de la représentation du raisonnement humain. Sur le plan technique [Herzig et al., 2017] argumente que les différentes logiques BDI ne répondent pas au *problème du décor* discuté précédemment ni au problème de la révision d'intention (*i.e* abandonner une intention de bas niveau de préférence à une intention de haut niveau). Du point de vue de la cognition humaine, [Arnaud et al., 2017a] argumentent sur le fait que les modèles BDI ne considèrent pas les apports des sciences humaines, avec notamment :

— Les jugements subconscients, c'est-à-dire les règles de raisonnement que l'agent utilise, mais ne croit pas explicitement. Par exemple, une personne peut croire qu'il n'est pas raciste, mais effectuer un jugement basé sur la couleur de peau d'une personne sans s'en rendre compte.

— la dissonance cognitive, c'est-à-dire le fait de maintenir deux états contradictoires dans ses croyances (*e.g* avoir l'intention de faire un régime et prendre une sucrerie).

De plus, les auteurs mettent en avant des limitations techniques avec :

— la mise à jour des croyances qui ne prennent pas en compte les biais humains ;
— l'engagement dans les intentions qui ne prend pas en compte les intentions difficiles à abandonner (*i.e* le biais d'engagement [Staw, 1997]) ;
— le changement de contexte, c'est-à-dire le fait de changer de désirs et d'intentions en fonction de l'image que nous voulons renvoyer (*e.g* j'ai l'intention de fuir un danger, mais devant les autres je montre que j'ai l'intention de rester pour me montrer courageux).

Nous pouvons voir à travers cette littérature que les logiques BDI sont très populaires pour représenter le raisonnement humain, du fait de leur expressivité. Pourtant, dans le même temps, elles souffrent à la fois de limitations techniques, mais aussi de limitations dans la prise en compte des biais cognitifs et plus largement des décisions incohérentes des humains.

III.1.4.d Conclusion

Représenter les croyances d'un agent en logique a été très étudié dans la littérature des logiques *doxastiques* avec deux logiques importantes en intelligence artificielle : la *Dynamic Epistemic Logic* et la *Beliefs, Desires and Intentions*. La DEL se concentre sur l'évolution des croyances d'un agent face à des événements dans un contexte multiagents, là où la BDI se concentre sur le concept d'intention d'un agent pour les choix d'action de l'agent. La BDI a l'avantage de reposer sur des concepts cognitifs qui sont faciles à se représenter pour un non-expert, ce qui explique en partie le développement de la prise en compte de facteurs humains comme les émotions dans ces logiques. Toutefois, les logiques BDI, en majorité, souffrent d'une complexité computationnelle plus élevée que la DEL, ce qui rend l'implémentation de ces logiques plus difficile. À l'exception de l'approche "base de données" de Shoham qui offre un cadre de formalisation beaucoup plus simple.

Au-delà de la complexité de ces logiques *doxastiques*, les travaux sur les facteurs humains dans les logiques BDI ne permettent pas de capturer toute la complexité de la cognition humaine, notamment sur les décisions parfois incohérentes des humains, en faisant des hypothèses trop fortes sur la rationalité de l'humain. Nous verrons toutefois section III.3 qu'il existe quelques solutions proposées dans la littérature utilisant la DEL et la BDI pour capturer les biais cognitifs, responsables en partie de ces décisions incohérentes (voir chapitre II).

En plus de représenter les croyances et le raisonnement d'un agent, les logiques *doxastiques* doivent répondre à trois problématiques différentes sur les changements des croyances que nous allons aborder dans la section suivante.

III.1.5 Le changement de croyance

En plus de la dynamique du monde qui prend en compte les changements dus aux actions de l'agent, le modèle doit prendre en compte la dynamique des

croyances de l'agent, c'est-à-dire comment les croyances de l'agent changent dans le temps en fonction des nouvelles informations perçues ou des actions effectuées. Ces problématiques liées à la dynamique des croyances sont regroupées dans la littérature sous le nom de *changement de croyance*. Trois changements peuvent être distingués : la *révision de croyance*, la *mise à jour* et l'*extrapolation de croyance*. Les deux dernières problématiques sont liées aux problématiques de la dynamique du monde (sous-section III.1.2) en traitant l'inertie des croyances de l'agent à la place de l'inertie du décor.

Nous allons dans cette section définir chacune de ces problématiques.

III.1.5.a La révision de croyance

La révision de croyance est le mécanisme qui permet de retrouver la cohérence des croyances d'un agent face à une nouvelle information contradictoire [Gärdenfors et al., 1995]. Par exemple, reprenons le problème YSP et considérons que l'agent a pour croyances :

$$\left\{ \begin{array}{l} \mathrm{alive}(0), \neg\, \mathrm{loaded}(0), \\ \mathrm{Load}(0), \mathrm{Load}(0) \rightarrow \mathrm{loaded}(1), \\ \mathrm{Shoot}(1), (\mathrm{Shoot}(1) \wedge \mathrm{loaded}(1)) \rightarrow \neg\, \mathrm{alive}(2) \end{array} \right\}$$

mais observe finalement que $\mathrm{alive}(2)$. L'agent est alors face à une contradiction en observant que la dinde est vivante alors que ses croyances lui permettent de déduire que la dinde est morte. L'agent doit donc réviser ses croyances, c'est-à-dire abandonner un ensemble de croyances pour que celles-ci soient cohérentes avec l'observation. Par exemple, l'agent peut abandonner la croyance $\mathrm{Load}(0) \rightarrow \mathrm{loaded}(1)$, c'est-à-dire que l'agent croit que finalement l'action $load$ a échoué. Une autre solution pour retrouver la cohérence dans les croyances de l'agent est d'abandonner $\mathrm{Load}(0) \rightarrow \mathrm{loaded}(1)$ mais aussi $\neg\, \mathrm{loaded}(0)$. Cette solution n'est pas satisfaisante, car il n'était pas nécessaire d'abandonner $\neg\, \mathrm{loaded}(0)$ puisque $\mathrm{Load}(0) \rightarrow \mathrm{loaded}(1)$ suffisait. Il est donc important de n'abandonner que les croyances nécessaires, c'est-à-dire effectuer une *révision de croyance minimale*.

Ce problème a été étudié par [Alchourrón et al., 1985] et a donné lieu à la théorie AGM (d'après le nom des chercheurs). Leur proposition est de définir trois opérateurs, l'expansion $(+)$ qui est l'ajout d'une croyance, la contraction $(\div)$ qui est la suppression d'une croyance et enfin la révision de croyance $(*)$ qui est l'ajout d'une croyance tout en restant cohérent. Les auteurs définissent un ensemble d'axiomes qui caractérisent un opérateur de contraction et de révision *minimale* des croyances. Nous notons $Cn(K)$ l'ensemble des conséquences logiques inférées à partir de K. Les huit axiomes d'AGM pour un ensemble de croyances K et un opérateur de révision $*$ sont :

(1) **La clôture** : $K*\varphi = Cn(K*\varphi)$ l'ensemble K après une révision est fermé (*i.e* l'ensemble révisé contient toutes les conséquences logiques inférées par $K * \varphi$).

(2) **L'inclusion** : $K * \varphi \subseteq K + \varphi$ l'ensemble K après une révision par φ est un sous ensemble de K auquel nous ajoutons φ

(3) **La vacuité** : si $\neg\varphi \notin Cn(K)$ alors $K + \varphi \subseteq K * \varphi$. Si la négation de φ n'est pas inférée par K alors la révision de K par φ correspond à K auquel nous ajoutons φ.

(4) **Le succès** : $\varphi \in K*\varphi$, l'ensemble K après une révision par φ doit contenir φ.

(5) **L'extensionnalité** : si $\varphi \leftrightarrow \psi \in Cn(\emptyset)$ alors $K * \varphi = K * \psi$, si K est révisé par deux propositions logiques équivalentes alors l'ensemble résultant de cette révision est équivalent dans les deux cas.

(6) **La cohérence** : si $\neg\varphi \notin Cn(\emptyset)$ alors $\bot \notin Cn(K*\varphi)$, l'ensemble K révisé par φ est cohérent, si φ est cohérent.

(7) **La super-expansion** : $K*(\varphi\wedge\psi) \subseteq (K*\varphi)+\psi$, un élément de $K*(\varphi\wedge\psi)$ est aussi un élément de l'ensemble K révisé par φ auquel nous ajoutons ψ.

(8) **La sous-expansion** : si $\neg\psi \notin (K * \varphi)$ alors $(K * \varphi) + \psi \subseteq K * (\varphi \wedge \psi)$, si la négation de ψ n'est pas inférée par la révision de K par ψ alors tout élément de K révisé par φ auquel nous ajoutons ψ est aussi un élément de K révisé par ψ et φ.

Ainsi, tout opérateur de révision de croyance qui se veut minimal doit respecter les axiomes (1-8). De plus, la littérature a montré par la *Levi identity* [Levi, 1977] que la révision minimale pouvait être définie par une contraction minimale (elle aussi axiomatisée dans les travaux d'AGM) :

$$K * \varphi = (K \div \neg\varphi) + \varphi$$

et par la *Harper identity* [Harper, 1976] qu'une contraction minimale pouvait être construite à partir d'une révision minimale :

$$K \div \varphi = K \cap (K * \neg\varphi)$$

La révision de croyance étant un domaine de recherche à part entière, il existe de trop nombreuses définitions d'opérateurs de révision de croyances respectant AGM ou des variations d'AGM (*i.e* des opérateurs respectant un sous ensemble des axiomes d'AGM) pour être exhaustif. Un tour d'horizon de ces opérateurs a été effectué par [Fermé et al., 2011]. Nous pouvons toutefois citer deux variations d'AGM qui jouent un rôle particulier dans cette thèse : la *shielded contraction* et la *screened revision*.

La *screened revision* par [Makinson, 1997] consiste à définir un ensemble C appelé noyau qui est immunisé à la révision de croyance. C'est-à-dire que si $*_{sc}$ est l'opérateur de *screened revision* alors $C \cap K \subseteq K *_{sc} \varphi$: tout élément de C en commun avec K ne doit pas être retiré de l'ensemble K révisé par une proposition. Cette révision permet notamment de considérer que du point de vue de l'agent, certaines croyances ne peuvent être abandonnées au vu de leur importance.

La *shielded contraction* par [Fermé et al., 2001] qui consiste à abandonner l'axiome (4) de succès dans AGM afin de permettre de ne pas accepter une nouvelle information φ. Cette révision permet notamment de considérer des agents qui ne considèrent pas qu'une nouvelle information à plus de valeur qu'une ancienne information.

La *révision de croyance* est donc un processus important dans le raisonnement de l'agent puisqu'il consiste, face à une contradiction, à déterminer les informations et croyances que l'agent décide de garder. Dans le contexte de cette thèse où nous essayons de comprendre une décision incohérente d'un agent, nous posons l'hypothèse qu'une erreur dans le processus de *révision de croyance* peut être une cause potentielle d'une erreur de décision. Nous ne sommes pas les premiers à formuler une telle hypothèse et nous aborderons des travaux allant dans ce sens section III.3.

III.1.5.b La mise à jour

Tout comme dans la dynamique du monde, il existe une inertie sur les croyances de l'agent. Si l'agent croit que son pistolet est chargé au temps 0 alors par défaut l'agent croira au temps suivant que son arme est toujours chargée sauf si le monde a changé. Ces changements peuvent être explicites (*i.e* par les effets des actions de l'agent) ou implicites (*i.e.* une observation du monde). La *mise à jour* est le mécanisme qui permet de prendre en compte les changements explicites là où l'*extrapolation des croyances* prend en compte les changements implicites.

La *mise à jour* est un mécanisme qui diffère de la *révision de croyance* dans le sens où la révision permet de prendre en compte une nouvelle information sur le monde ou une information incohérente par rapport aux croyances. Au contraire, la *mise à jour* prend en compte une information qui a changé dans le temps du fait des actions de l'agent. Reprenons un exemple inspiré de YSP avec un agent ayant pour croyance au temps 1 :

$$\left\{ \begin{array}{l} \mathrm{alive}(0), \neg\, \mathrm{loaded}(0), \\ \mathrm{Load}(0), \mathrm{Load}(0) \rightarrow \mathrm{loaded}(1) \end{array} \right\}$$

La *mise à jour* correspond au fait de prendre en compte que la proposition $loaded$ change de valeur de vérité entre le temps 0 et 1 du fait de l'action $\mathrm{Load}(0)$ et que toutes les autres croyances gardent leur inertie. La *mise à jour* est donc liée au *problème du décor* dans le sens où ce mécanisme cherche aussi à ne changer que les propositions dues aux actions de l'agent.

Tout comme la *révision de croyance*, la littérature propose une axiomatisation de l'opération de *mise à jour minimale*, dit KM, inspirée de AGM avec les travaux de [Katsuno et al., 1992]. Là encore, comme la littérature de la *révision de croyance*, la littérature de *la mise à jour* contient de nombreuses variations afin de palier les limites d'une *mise à jour minimale*, par exemple [Herzig et al., 1999, Doherty et al., 1998].

Le processus de *mise à jour* est aussi important que le processus de *révision de croyance* en prenant en compte les informations du monde qui changent en fonction des actions de l'agent. Là encore, nous pouvons poser l'hypothèse qu'une erreur dans le processus de *mise à jour* pourrait expliquer des décisions incohérentes. Par exemple, un agent qui effectue une action en oubliant certaines conséquences de cette action.

III.1.5.c L'extrapolation des croyances

L'*extrapolation des croyances* prend en compte les changements implicites des croyances, c'est-à-dire les changements dans le temps qui proviennent des observations du monde [Dupin de Saint-Cyr et al., 2011]. Prenons encore un exemple inspiré de YSP avec un agent ayant pour croyance au temps 2 :

$$\left\{ \begin{array}{l} \mathrm{alive}(0), \neg\, \mathrm{loaded}(0), \\ \mathrm{Load}(0), \mathrm{Load}(0) \rightarrow \mathrm{loaded}(1) \\ \mathrm{Wait}(1) \end{array} \right\}$$

et observe au temps 2 que son pistolet a été déchargé entre-temps : $\neg\, \mathrm{loaded}(2)$. Avec l'inertie des croyances, l'agent devait conclure naturellement que l'état de la proposition $loaded$ n'a pas changé après avoir fait l'action $Load$. Toutefois, le monde a changé entre temps avec le déchargement de l'arme, l'*extrapolation des croyances* consiste alors à prendre en compte que la proposition $loaded$ doit changer de valeur. [Dupin de Saint-Cyr et al., 2011] montre que l'*extrapolation de croyances* est finalement une instance d'une *révision de croyance* sur les clauses d'inertie d'une logique indexée temporellement. Un opérateur de type AGM peut être alors utilisé pour résoudre l'*extrapolation des croyances*. Nous utiliserons cette propriété dans cette thèse pour calculer ce processus.

L'*extrapolation de croyance* est, tout comme les deux processus de changement de croyances précédents, important pour comprendre l'évolution des croyances d'un agent. Le monde peut changer sans l'intervention de l'agent et celui-ci doit le prendre en compte. Là encore une erreur qui intervient dans le processus de d'*extrapolation de croyance* pourrait potentiellement expliquer une erreur de décision. Par exemple, un agent qui prend une décision basée sur des anciennes informations en ne prenant pas en compte de nouvelles informations qui lui sont disponibles.

III.1.5.d Conclusion

Nous pouvons voir que la littérature met en avant trois opérations importantes pour la formalisation des changements de croyance dans le temps d'un agent, à savoir la *révision de croyance*, la *mise à jour* et l'*extrapolation des croyances*. Toutes ces opérations sont nécessaires pour transiter d'un état de croyance à un autre et

sont attendues dans une logique cherchant à formaliser les états de croyances d'un agent.

Dans le contexte de cette thèse sur l'analyse de décision incohérente, nous pensons que certaines décisions erronées peuvent être expliquées par une erreur dans la transition entre deux états de croyances. Par conséquent, une erreur pourrait survenir dans l'un des trois processus de changement de croyance ou pourquoi pas plusieurs erreurs sur plusieurs processus en même temps.

III.1.6 Conclusion

Représenter les actions et croyances d'un agent dans un environnement dynamique est un challenge complexe. Il est nécessaire de prendre en compte de nombreuses problématiques liées au changement du monde dû aux actions des agents, à savoir le *problème du décor*, le *problème de la qualification*, le *problème de la ramification*. À cela s'ajoutent les problèmes liés aux changements des croyances dans le temps de l'agent avec la *révision de croyances*, la *mise à jour* et l'*extrapolation de croyances*, tout en considérant dans notre contexte que des erreurs peuvent survenir dans ces changements de croyances.

La littérature sur ces problématiques qui remontent au début de l'intelligence artificielle est vaste et propose de nombreuses solutions permettant de répondre à chacune des problématiques. Ces solutions consistent généralement à proposer des opérateurs pour gérer l'aspect dynamique du modèle (changement des croyances, du monde) et des opérateurs pour représenter l'état épistémique de l'agent (croyances, désirs, intentions, connaissances, etc). Une des approches les plus populaires pour représenter de manière logique un agent est la logique BDI qui repose sur des concepts faciles à représenter pour un non spécialiste. De ce point de vue, ce type de logique représente un candidat de choix pour notre problématique d'explication d'une décision erronée d'un humain. Toutefois, ces logiques souffrent de limitations techniques en ne répondant pas forcément à toutes les problématiques sur la dynamique du monde et des croyances. De plus, une limitation majeure par rapport à la problématique de cette thèse est la tendance des logiques BDI à oublier les apports des sciences humaines et notamment la prise en compte des biais cognitifs dans le raisonnement humain. La littérature nous offre donc un début de solution pour représenter un agent dans un environnement dynamique. Reste la question du diagnostic du raisonnement d'un tel agent lors d'une décision erronée ainsi que la prise en compte des biais cognitifs dans le raisonnement humain. Nous aborderons la littérature sur ces deux problématiques dans les sections suivantes.

III.2 Le diagnostic en intelligence artificielle

Une des problématiques de cette thèse est de diagnostiquer la prise de décision erronée d'un agent dans un environnement dynamique. Nous avons vu section III.1 que la littérature sur la formalisation logique des croyances et des actions d'un agent offre déjà de nombreux outils pour répondre à la problématique de la représentation du raisonnement d'un agent dans un environnement dynamique. Il reste alors la question de trouver une explication à une situation où un agent prend une décision erronée. Le sous-domaine de l'intelligence artificielle s'intéressant à cette problématique est le diagnostic. Nous allons voir dans cette section que tout comme la représentation des croyances et des actions, la littérature sur ce domaine remonte au début de l'intelligence artificielle et n'est pas une nouvelle problématique.

Le but général du diagnostic est de détecter et de trouver une explication au comportement anormal d'un système. Ce but général peut s'inscrire dans des contextes différents et cacher finalement des problématiques différentes. Nous allons voir dans cette section que trois approches répondant à trois problématiques différentes de diagnostic se distinguent dans la littérature : l'approche par déduction, par abduction et par cohérence. Les deux dernières approches étant regroupées dans la littérature sous le nom de *model-based diagnosis*. Nous allons présenter chacune de ces approches avec leur principe général, un exemple et les avantages et limites de l'approche. Nous verrons ensuite le lien très fort entre le diagnostic et les opérateurs de révision de croyance.

III.2.1 Approche par déduction

Principe Historiquement, les premières solutions proposées dans le domaine du diagnostic en intelligence artificielle sont les *systèmes experts* s'intéressant à la problématique de simuler le raisonnement d'un expert pour diagnostiquer un système [Buchanan et al., 1984]. Ces approches reposent sur deux ensembles :

- L'ensemble des règles de diagnostic de l'expert $\mathcal{R}$ qui doivent être utilisées pour déduire une explication à partir des observations du système.
- Les observations OBS qui sont un ensemble de propositions qui représente les observations du système à expliquer.

Les règles de $\mathcal{R}$ suivent le schéma $effets \rightarrow causes$, afin de modéliser le raisonnement de l'expert pour diagnostiquer le système. Un diagnostic Δ possible des observations OBS est alors l'ensemble des conséquences logiques des règles $\mathcal{R}$ à partir de OBS.

Exemple Considérons un exemple inspiré de YSP où nous souhaitons diagnostiquer avec une approche par déduction la situation où l'agent appuie sur la détente du pistolet, mais ne tue pas la dinde. Nous faisons l'hypothèse que dans un tel cas, nous avons la connaissance que les explications possibles sont que le pistolet utilisé par l'agent peut s'enrayer (jammed) ou que le chargeur du pistolet soit vide

(empty). Nous avons alors :

$$\mathcal{R} \equiv \left\{ \begin{array}{l} \text{alive}(3) \rightarrow (\neg \text{Shoot}(2) \vee \text{jammed}(2) \vee \neg \text{loaded}(2)) \\ \neg \text{loaded}(2) \rightarrow (\neg \text{Load}(1) \vee \text{empty}(1)) \end{array} \right\}$$

$$OBS \equiv \{\text{alive}(3), \text{Shoot}(2), \text{Load}(1)\}$$

Les diagnostics possibles Δ sont alors :

$$\Delta_1 \equiv \{\text{jammed}(2)\}$$

$$\Delta_2 \equiv \{\text{empty}(1)\}$$

Limites Cette approche a l'avantage de reposer sur un mécanisme simple de la logique, la déduction. Néanmoins, cette approche a de nombreuses limitations. par exemple si le système expert ne trouve pas de règle à appliquer pour une observation alors aucun diagnostic ne sera retourné. De ce fait, il n'existe pas d'explication possible pour un comportement qui n'est pas connu des experts. De plus, la construction de ces systèmes de manière empirique n'est pas possible pour certains domaines où l'erreur d'un système est trop coûteuse (*e.g* le domaine de l'aviation, du nucléaire, etc). C'est pourquoi le domaine du diagnostic s'est tourné vers d'autres solutions orientées modèles (*model-based diagnosis*) qui ne viennent pas décrire le raisonnement d'un diagnostic d'un expert du système, mais décrivent les connaissances sur le fonctionnement du système étudié (*e.g* les interactions entre les composants du système). Deux paradigmes existent dans la littérature reposant sur le principe de modèle : l'approche par abduction et l'approche par cohérence. Nous allons dans les deux prochaines sections définir chacune de ces approches.

III.2.2 Approche par abduction

Principe L'approche par abduction se place généralement sur la problématique suivante [Poole, 1994] :

Il existe des connaissances sur les erreurs et symptômes de ces erreurs. L'objectif est de déduire les observations du système à partir de ces connaissances.

Le modèle le plus influent de la littérature pour le diagnostic par abduction est le modèle dit *Theorist* de [Poole et al., 1987]. Cette approche repose sur trois ensembles :

- les faits $\mathcal{F}$ qui sont un ensemble de formules et propositions logiques qui sont vraies dans le monde et cohérentes logiquement.
- les hypothèses $\mathcal{H}$ qui sont un ensemble de formules et propositions logiques qui sont acceptées comme une partie de l'explication possible pour les observations sur le système.

— les observations OBS qui sont un ensemble de propositions qui représente les observations du système à expliquer.

De ces trois ensembles, Δ est un diagnostic tel que :

— $\Delta \subset \mathcal{H}$, le diagnostic est un sous ensemble des explications possibles.

— $(\mathcal{F} \cup \Delta) \models OBS$, les faits et le diagnostic permettent de déduire les observations.

— $(\mathcal{F} \cup \Delta) \not\models \bot$, les faits et le diagnostic sont cohérents.

Chaque règle logique de chaque ensemble suit le schéma *causes* $\rightarrow$ *effets* à l'inverse des systèmes experts qui suivent le schéma *effets* $\rightarrow$ *causes*. Cette inversion de paradigme permet aux approches par abduction de rester dans la logique de modéliser le comportement du système étudié et non le raisonnement d'un expert qui effectue un diagnostic sur ce système.

Exemple Reprenons l'exemple inspiré du problème YSP vu sous-section III.2.1 et considérons que nous souhaitons diagnostiquer avec une approche par abduction. Nous faisons l'hypothèse que nous avons la connaissance que le pistolet utilisé par l'agent peut parfois s'enrayer ou que le chargeur du pistolet soit vide.

$$\mathcal{F} \equiv \{ \text{Load}(1), \text{Shoot}(2) \}$$

$$\mathcal{H} \equiv \left\{ \begin{array}{l} \text{empty}(1), \text{jammed}(2) \\ (\text{Load}(1) \wedge \text{empty}(1)) \rightarrow \neg\, \text{loaded}(2) \\ (\text{Shoot}(2) \wedge \neg\, \text{loaded}(2)) \rightarrow \text{alive}(3) \\ (\text{Shoot}(2) \wedge \text{loaded}(2) \wedge \text{jammed}(2)) \rightarrow \text{alive}(3) \end{array} \right\}$$

$$OBS \equiv \{\text{alive}(3)\}$$

Ici les faits $\mathcal{F}$ décrivent que nous savons que l'agent a chargé le pistolet et pressé la détente du pistolet. Les hypothèses $\mathcal{H}$ définissent les différentes explications possibles avec le chargeur vide et le pistolet enrayé. Les observations OBS définissent la proposition à expliquer, à savoir que la dinde est toujours vivante. Nous avons comme diagnostic Δ possible par abduction :

$$\Delta_1 \equiv \left\{ \begin{array}{l} \text{empty}(1) \\ (\text{Load}(1) \wedge \text{empty}(1)) \rightarrow \neg\, \text{loaded}(2) \\ (\text{Shoot}(2) \wedge \neg\, \text{loaded}(2)) \rightarrow \text{alive}(3) \end{array} \right\}$$

$$\Delta_2 \equiv \left\{ \begin{array}{l} \text{jammed}(2) \\ (\text{Shoot}(2) \wedge \text{loaded}(2) \wedge \text{jammed}(2)) \rightarrow \text{alive}(3) \end{array} \right\}$$

Limites Cette approche a l'avantage de proposer une modélisation du problème de diagnostic plus naturelle que l'approche par déduction. Il est plus facile de représenter le fonctionnement d'un système que de représenter le raisonnement de diagnostic d'un expert sur ce système. Toutefois, cette approche ne résout pas un inconvénient majeur de l'approche précédente, la nécessité de connaître ou poser

des hypothèses sur les erreurs et symptômes du système. La dernière approche par cohérence vient répondre à cette problématique.

III.2.3 Approche par cohérence

Principe Les diagnostics basés sur la cohérence (*consistency based diagnosis*) se placent généralement sur la problématique suivante [Poole, 1994] :

Il existe des connaissances sur comment le système fonctionne normalement. Il n'existe pas de connaissance sur les erreurs et symptômes de ces erreurs. L'objectif est d'identifier les déviations du comportement attendu du système.

Le *consistency based diagnosis* proposé par [Reiter, 1987] est un modèle logique pour le diagnostic automatique qui consiste à retrouver la cohérence entre la description du comportement du système et les observations. Ce modèle comporte trois ensembles logiques :

- SD un ensemble de formules logiques qui décrit le système,
- ASS un ensemble de prédicats qui décrit les *hypothèses* de la forme $\neg ab(c)$, *i.e* le composant c est supposé se comporter normalement,
- OBS une conjonction de prédicats qui décrit une observation du système.

Quand $SD \cup ASS \cup OBS$ est incohérent, un diagnostic Δ est un ensemble minimal d'*hypothèses* tel que $SD \cup (ASS \setminus \Delta) \cup OBS$ est cohérent. En d'autres termes, un diagnostic est un ensemble minimal d'éléments dont on doit supposer qu'ils ont un comportement anormal pour retrouver la cohérence avec les observations.

Exemple Considérons le même exemple que sous-section III.2.2 avec une approche par cohérence et sachant que nous n'avons pas la connaissance sur l'enrayement possible du pistolet ou du chargeur vide. Nous faisons l'hypothèse que le pistolet et le chargeur sont fonctionnels :

$$
\begin{aligned}
ASS &\equiv \{\neg ab(gun), \neg ab(load)\} \\
SD &\equiv \left\{ \begin{array}{l} \mathrm{Load}(1) \wedge \neg ab(load) \rightarrow \mathrm{loaded}(2) \\ (\mathrm{Shoot}(2) \wedge \mathrm{loaded}(2) \wedge \neg ab(gun)) \rightarrow \neg\, \mathrm{alive}(3) \end{array} \right\} \\
OBS &\equiv \{\mathrm{Load}(1), \mathrm{Shoot}(2), \mathrm{alive}(3)\}
\end{aligned}
$$

Nous trouvons alors comme diagnostic Δ possible :

$$
\begin{aligned}
\Delta_1 &\equiv \{\neg ab(gun)\} \\
\Delta_2 &\equiv \{\neg ab(load)\}
\end{aligned}
$$

Limites Cette approche a l'avantage de ne pas reposer sur des explications préétablies par le modélisateur du système et permet de déterminer les composants du système possiblement responsables d'une erreur. Toutefois, l'inconvénient est

que les solutions proposées par l'approche par cohérence sont moins précises que les autres approches. En effet, elles permettent de déterminer les composants du système défectueux (*i.e* ce qui ne va pas), mais pas la raison sous-jacente qui a mené à ce système défectueux (*i.e* ce qui s'est passé).

III.2.4 Relation entre la révision de croyance et les model-based diagnostic

La littérature des diagnostics basés sur les modèles met en avant à travers plusieurs travaux le lien entre ce type de diagnostic et la révision de croyance. Si nous reprenons les notations de la sous-section III.2.3, [Reiter, 1987] définit la construction du diagnostic par :

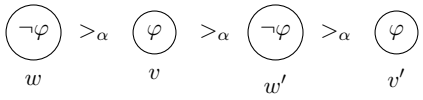
$\Delta \subseteq ASS$ est un diagnostic pour (SD, ASS, OBS) si et seulement si Δ est un hitting set pour l'ensemble des conflict set minimaux de (SD, ASS, OBS).

où un *hitting set* pour une collection d'ensembles est un ensemble dont l'intersection avec chacun des ensembles de la collection est non vide. Un *conflict set* C est un sous ensemble de ASS tel que $SD \cup OBS$ est incohérent. [Wassermann, 2000] montre que cette construction est équivalente à calculer une révision de croyances respectant AGM (présenté au III.1.5.a). Ainsi, un problème de diagnostic peut être traduit en un problème de révision de croyance et l'algorithme de Reiter pour calculer un diagnostic peut être utilisé pour calculer une révision de croyances.

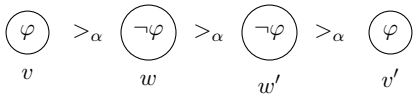
Avant les travaux de Wassermann, [Boutilier et al., 1995] montraient déjà ce lien entre la révision de croyances et les *model-based* diagnostic en utilisant un modèle de plausibilité. Ce modèle permet de représenter le fait qu'un agent considère plusieurs mondes possibles en considérant que certains sont plus plausibles que d'autres. Ainsi, un modèle de plausibilité $M = (W, \geq, V)$ est composé :

- un ensemble de mondes non vide W qui représente tous les états possibles ;
- une fonction $\geq$ qui attribue pour chaque agent $\alpha \in \mathcal{A}$ une relation de plausibilité $\geq_\alpha$ sur les mondes W, où $w \geq_\alpha v$ indique que pour l'agent α, le monde w est plus plausible que le monde v ;
- une fonction d'évaluation V qui attribue à chaque monde les propositions qui y sont vraies.

Dans le cas d'une *révision de croyance* par φ, les mondes les plus plausibles sont initialement ceux qui contiennent $\neg\varphi$. L'idée est alors qu'après une *révision de croyance*, les mondes contenants φ deviennent plus plausibles que ceux qui contiennent $\neg\varphi$. À partir de cette idée, les auteurs définissent la *conservative upgrade* qui correspond à promouvoir seulement les mondes les plus plausibles contenants φ et ne pas bouger les autres. Par exemple, si nous avons les plausibilités suivantes pour l'agent α pour chaque monde possible :



Après une *conservative upgrade* pour φ nous avons comme relation de plausibilité :



Le monde v devient le plus plausible, car φ y est vrai et il est le plus plausible des mondes où φ est vrai. À partir de cette opération, les auteurs montrent qu'ils arrivent à reconstruire les deux approches types *model-based diagnosis*, à savoir l'approche par abduction et par cohérence. Dans le même temps, les auteurs montrent le lien entre l'approche par abduction et par cohérence en démontrant que dans le cas où il n'y a pas de connaissance complète sur les erreurs possibles du système, alors les deux approches coïncideront.

Nous pouvons donc voir que les opérateurs de révision de croyances représentent un outil non pas seulement pour la gestion des croyances d'un agent intelligent, mais aussi pour le diagnostic d'un système. Notre problématique de thèse touchant à la fois ces deux domaines, la littérature montre que l'opération de révision de croyance ne peut être qu'un élément central de notre approche.

III.2.5 Conclusion

Nous pouvons voir qu'il existe plusieurs approches possibles pour un problème de diagnostic, les *model-based* diagnostic étant les plus populaires en intelligence artificielle. Ainsi, de nombreux travaux ont enrichi ces approches avec par exemple l'introduction du temps [Brusoni et al., 1998] ou des probabilités [Lucas, 2001]. Chacune de ces approches a ses avantages et inconvénients et le choix d'une application d'une de ces deux approches repose essentiellement sur la problématique que nous cherchons à résoudre. Dans le cas où les comportements anormaux d'un système sont connus, il est préférable de se tourner vers l'approche par abduction. Cependant, s'il y a peu d'expérience sur les défauts du système, alors il est préférable de se tourner vers une approche basée sur la cohérence. Dans cette thèse, les défauts que nous cherchons à diagnostiquer sont des biais cognitifs. Nous avons pu voir section II.3 que la littérature des sciences humaines n'a pas de consensus sur la classification des biais et sur les caractéristiques de ces biais. De ce constat, il semble difficile d'écrire des règles décrivant les comportements anormaux qui résultent des biais. Nous nous retrouvons alors sur une problématique à la *consistency-based diagnosis*, c'est-à-dire qu'il existe des connaissances sur la manière dont le système fonctionne normalement, mais des connaissances floues sur les erreurs de ce système. Nous pensons que dans notre approche, il est préférable de détecter les déviations du comportement attendu et à partir de là de déterminer les biais cognitifs.

Enfin, la littérature sur le diagnostic en intelligence artificielle montre qu'un opérateur de révision de croyance permet de calculer un diagnostic. Cette opération semble donc nécessaire pour à la fois gérer les croyances de l'agent modélisé,

mais aussi pour diagnostiquer la solution. Nos travaux doivent donc donner une importance toute particulière à ces opérateurs de révision.

III.3 Les biais cognitifs dans un cadre formel

Nous avons vu au chapitre II la nécessité de capturer les biais cognitifs pour comprendre les erreurs humaines. La littérature présente des modèles formels très divers pour cet objectif : automate, architecture BDI, révision de croyance et DEL. Nous allons pour chaque modèle présenter ses avantages et inconvénients.

III.3.1 Automate à états finis

Pour capturer les biais cognitifs de confirmation et de conformisme dans une campagne de vaccination, [Voinson et al., 2015] utilise un automate à états finis. Pour cela, chaque état de l'automate représente à la fois :

— le statut épidémiologique de chaque agent (*i.e* exposé à l'infection, infecté, soigné après vaccination, soigné naturellement) ;
— l'opinion de chaque agent sur la vaccination (positive ou négative).

La transition entre les états épidémiologiques se fait en fonction du taux de transmission, d'infection β et du taux de vaccination des individus θ. À cela s'ajoute deux autres états qui indiquent le nombre d'individus qui souffrent d'effets secondaires de la vaccination C^V et de l'infection C^I. Les biais cognitifs sont alors représentés formellement par les transitions entre les états d'opinions (Ω pour positif à négatif et Ω' pour négatif à positif). En effet, pour représenter le biais de confirmation, les auteurs considèrent qu'un agent avec une opinion positive va donner plus de poids au coût de l'infection C^I alors qu'au contraire un agent avec une opinion négative va donner plus de poids au coût de vaccination C^V. De plus, pour représenter les biais de conformisme, les transitions entre les états d'opinions prennent en compte l'opinion majoritaire dans l'ensemble de la population, que les agents auront tendance à suivre.

Ce modèle a l'avantage d'être simple et permet de comprendre et de quantifier les mécanismes qui ont poussé les agents à une décision qui n'est pas attendue (ne pas se vacciner). Toutefois, ce modèle présente plusieurs limitations. La première limitation est que ce modèle ne peut être appliqué que dans un contexte de vaccination. En effet, les croyances des agents se résument à des opinions sur la vaccination, ce qui empêche une généralisation à d'autres prises de décisions. Enfin, une seconde limitation liée à la première est que la prise en compte d'autres biais cognitifs plus complexes, comme les corrélations illusoires entre un évènement négatif et la campagne de vaccination malgré l'indépendance des deux évènements, n'est pas abordée.

III.3.2 Architecture BDI

Dans un autre domaine d'application (situation de feux de forêts) et en utilisant un modèle différent, [Arnaud et al., 2017b] cherchent à capturer trois biais cognitifs (négligence des probabilités, effet Semmelweis et effet de vérité illusoire). Pour cela, les auteurs se basent sur une architecture BDI où chaque croyance est associée à une probabilité de 0 à 100, où 0 représente le fait que l'agent croit que c'est certainement faux, 100 que c'est certainement vrai et 50 qu'il est incertain. Les biais cognitifs sont alors représentés par des fonctions mettant à jour ces probabilités en fonction des informations reçues. Par exemple, dans le cas de la vérité illusoire, plus l'information est répétée à l'agent, plus la probabilité de la croyance correspondante augmente.

Ce modèle à l'avantage d'avoir un modèle d'agent très expressif en se basant sur une architecture BDI. De plus, la construction des biais cognitifs comme une fonction probabiliste permet de facilement ajouter et prendre en compte de nouveaux biais, tout en ayant une évaluation quantitative de l'impact des biais. Toutefois, les auteurs ne précisent pas certains termes utilisés dans le calcul de la probabilité d'une croyance comme "*is small*", "*not perceive to be dire*", "*perceived to be extremely favourable*", etc., qui n'ont pas une quantification précise ou n'ont pas une fonction permettant de retourner la vérité de ces termes. Enfin, il n'est pas clair comment les différents biais implémentés s'agencent entre eux. Par exemple, si nous ajoutons une fonction représentant le *biais de confirmation* qui va augmenter la probabilité d'une croyance si une information va dans le sens de cette croyance, est-ce que le biais de *vérité illusoire* va être aussi déclenché ou un des deux biais est-il prioritaire sur l'autre ?

III.3.3 Révision de croyance

[Dutilh Novaes et al., 2016] compare la logique préférentielle et la révision de croyance pour capturer le *biais de croyance*, c'est-à-dire la tendance à juger des arguments en fonction de la plausibilité de la conclusion et non à quels points ils supportent la conclusion. Par exemple, si nous avons les propositions :

Certaines actrices ne sont pas belles
Toutes les femmes sont belles

alors nous devons conclure que "certaines actrices ne sont pas des femmes" car nous avons le système logique : "Certains A ne sont pas B. Tous les C sont B. Par conséquent, certains A ne sont pas C". Or une majorité de personnes affirme que cette conclusion est incorrecte, car peu plausible, tombant dans un *biais de croyance*. Les auteurs étudient empiriquement sur différentes tâches liées au problème du *biais de croyance* les réponses des différents participants et évaluent quels modèles entre la logique préférentielle et la *screened revision* capture le mieux les réponses des participants. Les auteurs en concluent que la logique préférentielle ne

capture pas le fait que les personnes ont tendance à ne pas tirer des conclusions si celles-ci rentrent en conflit avec des anciennes croyances. Toutefois, la *screened revision*, en définissant les anciennes croyances comme le noyau immunisé à la révision, permet de capturer le phénomène du *biais de croyance*.

Ces travaux permettent de montrer empiriquement que les opérateurs de révision de croyance ont du potentiel pour capturer certains biais cognitifs. Toutefois, les travaux des auteurs ne s'intéressent qu'à un seul biais et sont uniquement liés à la logique du raisonnement. Il est possible que sur des biais prenant en compte d'autres facteurs (*e.g* les émotions), les opérateurs de révision de croyance ne suffisent pas à capturer le phénomène ou ne capturent qu'une partie du phénomène, comme l'ont démontré les auteurs pour la logique préférentielle.

III.3.4 Dynamic Epistemic Logic

[Solaki et al., 2021] propose de s'inspirer du *dual-process model* [Frankish, 2010] en se basant sur une Dynamic Epistemic Logic. Pour cela, les auteurs introduisent la notion de *monde impossible* qui correspond à des mondes (*i.e* un ensemble de propositions vraies) qui ne sont pas clos sous conséquence logique. Par exemple :

$\{s, s \to p, p\}$	est un *monde possible*
$\{s, s \to p\}$	est un *monde impossible*

Le modèle des auteurs est composé alors d'une relation de plausibilité entre l'ensemble des mondes possibles et impossibles, un coût cognitif pour chaque règle de raisonnement et une capacité cognitive pour l'agent. Par exemple si l'agent à une capacité cognitive de 1 et que la règle de raisonnement $s \to p$ est de coût 2, alors l'agent ne peut pas l'appliquer et tombe dans le *monde impossible* $\{s, s \to p\}$.

Enfin, le langage du modèle est une Dynamic Epistemic Logic avec la particularité que l'opérateur de changement de modèle $[\alpha]\varphi$ est de deux types afin de représenter le *dual-process model* :

(*System 1*) $[\Psi \Uparrow]\varphi$ représente un changement de croyance rapide qui se lit "après avoir mis à jour avec Ψ, φ est vrai" où Ψ est une formule propositionnelle. Cela équivaut à ce que les mondes satisfaisants Ψ deviennent plus plausibles que les mondes satisfaisants $\neg\Psi$.

(*System 2*) $[R_k]\varphi$ représente l'application d'une règle de raisonnement et se lit "après l'application de la règle R_k, φ est vrai". Cela représente un changement plus coûteux cognitivement pour l'agent.

Le principe du modèle est alors en deux phases lorsque l'agent reçoit une information ψ alors que $\neg B\psi$:

(1) Dans la première phase, l'agent prend en compte ψ par l'opérateur $[\psi \Uparrow]B\psi$. Les mondes impossibles satisfaisants ψ ont leur plausibilité augmentée.

(2) Dans une deuxième phase, pour chacune des règles de raisonnement des mondes les plus plausibles et si le coût cognitif en fonction de la capacité le permet, les règles sont appliquées. Les mondes résultant de l'application de ces règles deviennent alors plus plausibles. Par exemple, si nous avons $R_1 = \psi \rightarrow \varphi$ dans un monde plausible, alors un monde satisfaisant $[R_1]\varphi$ devient plus plausible.

Ainsi, la construction de l'état de croyance de l'agent se fait d'abord par l'incorporation d'une information et à partir de là l'application de règles de raisonnement en fonction des limitations cognitives de l'agent. Par conséquent, l'agent n'est pas omniscient et peut avoir des croyances biaisées du fait qu'il n'a pas accès à toutes les conclusions.

Ce modèle a l'avantage de se baser sur l'une des théories les plus utilisées pour expliquer les biais (*i.e dual-process model*). De plus, le modèle se base sur des outils connus des modèles formels la Dynamic Epistemic Logic ce qui permet d'ajouter de nouveaux composants au modèle facilement. Toutefois, les auteurs ont fait le choix d'un modèle descriptif et réaliste en faisant beaucoup de concessions sur la simplicité du modèle. Il en résulte un modèle complexe en termes de calcul, notamment sur la génération des mondes impossibles, qui est difficilement implémentable. Il reste cependant le fait que se rapprocher d'une telle modélisation permet de capturer de nombreux biais cognitifs.

III.3.5 Conclusion

Au vu des différentes solutions proposées, nous faisons le même constat que [Solaki et al., 2021]. Prendre en compte les biais cognitifs dans un cadre formel est un compromis entre un modèle simple qui ne capture pas l'ensemble des biais ou un modèle complexe dont le coût computationnel est trop élevé. Nous notons aussi que les mécanismes de changement de croyance et de représentation formelle d'un agent offrent tous deux un moyen de formaliser les biais cognitifs (révision de croyance, BDI, DEL), ce qui pose la question de la complémentarité de ces approches pour la formalisation des biais. Enfin, nous notons que toutes les approches citées ci-dessus sont prédictives dans le sens où elles cherchent à déterminer les croyances d'un agent biaisé face à une situation. Dans cette thèse, nous adoptons la démarche inverse. Nous cherchons une approche de diagnostic où, étant donné une situation et une prise de décision de l'agent, on cherche à déterminer quel biais permet d'expliquer sa décision.

III.4 Positionnement

Les travaux en sciences humaines sur la question de l'erreur humaine que nous avons abordés au chapitre II mettent en avant deux phases importantes pour le diagnostic complet d'une erreur humaine. La première phase consiste à déterminer

les croyances possibles de l'agent qui sont cohérentes avec la décision incohérente du point de vue de l'agent. Cela permet de donner une cause à la décision de l'agent : "l'agent a fait a car il croyait φ". La deuxième phase consiste à comprendre la raison de la croyance erronée de l'agent en se reposant sur une explication à base de biais cognitifs : "L'agent croyait φ car il est tombé dans le biais ψ". Chaque biais offrant une explication possible à une croyance erronée. Il nous semble donc naturel qu'un modèle de diagnostic d'erreur humaine suive cette même idée avec deux modèles indépendants : un modèle qui recherche les croyances possibles de l'agent et un modèle qui cherche une raison à ces croyances via les biais cognitifs.

Dans cette même littérature de l'étude de l'erreur humaine, nous pouvons voir qu'il n'existe pas de classification et de caractérisation des biais cognitifs qui fasse consensus. De plus, les différentes classifications proposées dans la littérature (notamment les explicatives qui sont plus proches de notre problématique) reposent sur des définitions peu adaptées à une représentation formelle. Il semble donc nécessaire de créer une taxonomie qui n'a pas encore été explorée par la littérature sur les biais : une taxonomie formelle. Une telle taxonomie aurait l'avantage de poser des concepts précis derrière les biais cognitifs pour offrir une base solide afin de formaliser chaque biais et donc une base pour le développement du deuxième modèle de notre approche.

Du côté de l'informatique, diagnostiquer un système logique est une question qui remonte au début de l'intelligence artificielle. Plusieurs approches sont possibles en fonction de la connaissance des erreurs que nous avons sur le système étudié. Dans notre cas, nous considérons que nous n'avons aucune connaissance précise sur les erreurs cognitives d'un humain. En effet, comme nous avons argumenté précédemment, les caractéristiques des biais cognitifs ne sont pas définies de manière formelle (il s'agit de description en langage naturel de propriétés cognitives) et ne permettent pas la construction de règles précises liant la cause à l'effet. C'est pourquoi le *consistency-based diagnosis* semble approprié dans notre problématique. De plus, celui-ci correspond tout à fait au raisonnement des enquêteurs décrit par [Dekker, 2006] (voir chapitre II), l'enquêteur recherche à retrouver la cohérence entre la décision et le point de vue de l'agent.

Toujours du côté de l'informatique, représenter un agent qui effectue des actions dans un environnement dynamique est une question, elle aussi, assez ancienne. De nombreux modèles logiques ont été développés pour s'attaquer aux différents problèmes de changements liés aux actions, observations et croyances. Dans notre cas, nous cherchons une logique permettant de répondre à ces problèmes de changements, mais aussi reposant sur des concepts proches de la cognition humaine afin d'offrir une explication compréhensible et exploitable pour représenter les biais cognitifs. En ce sens, les logiques BDI et leurs trois concepts centraux de croyances, intentions et désirs sont les logiques qui offrent une piste intéressante, car proche de la cognition humaine. Au sein de cette famille, nous pensons que l'approche à la base de données de [Shoham, 2009] qui cherche la cohérence entre la base de

croyances et la base d'intentions (d'actions) de l'agent se rapproche de la méthode décrite par Dekker (recherche de croyance cohérente avec la décision de l'agent). Cette logique nous semble donc parfaitement adaptée à cette problématique.

De plus, les quelques travaux en informatique cherchant à formaliser les biais pour un comportement non attendu le font dans un contexte de prédiction et non de diagnostic. C'est-à-dire qu'ils offrent un modèle qui permet, à partir des croyances de l'agent, de prédire la décision biaisée de celui-ci. Au contraire, nous cherchons, à partir de la décision biaisée d'un agent, à retrouver les croyances biaisées de l'agent. De plus, ces travaux mettent en avant que la représentation des biais repose sur un compromis entre la simplicité du modèle, qui permet une expérimentation et une validation du modèle, mais capture peu de biais, et la complexité du modèle, qui permet de capturer plus de biais, mais laisse peu de place à l'expérimentation et la validation. Il sera donc nécessaire dans ces travaux de thèse de trouver un juste équilibre entre simplicité et complexité.

Enfin, nous souhaitons conclure sur un point qui semble important sur l'ensemble de cette littérature. Il existe un lien non négligeable entre le diagnostic d'un système logique, la représentation d'un agent dans un environnement dynamique en logique et la représentation des biais en logique. En effet, la recherche d'une solution minimale avec l'opérateur de révision de croyance à la AGM permet de répondre au problème du diagnostic et à certains problèmes liés aux changements dans un monde dynamique. De plus, [Dutilh Novaes et al., 2016] montrent que ce même opérateur est une piste encourageante pour capturer certains biais cognitifs. Nous pensons donc que cet opérateur ne peut être que central à notre approche du fait de son importance sur l'ensemble des problématiques que nous abordons dans cette thèse.

IV - Présentation de l'approche

Nous avons présenté au chapitre II une vue d'ensemble de l'étude de l'erreur humaine dans le domaine de la sécurité et des sciences cognitives, puis au chapitre III, nous avons fait un tour d'horizon des cadres formels en logique pour représenter les croyances et le raisonnement humain ainsi que les approches de diagnostic en logique. Ce travail bibliographique nous a permis de nous éclairer sur la démarche à suivre pour la construction de notre approche du diagnostic d'une décision erronée d'un agent. Pour illustrer cette approche, nous allons dans un premier temps détailler notre cas d'étude principal dans cette thèse : l'accident du vol d'Air France 447 entre Rio et Paris en 2009. Enfin, nous développerons l'approche générale qui s'inspire du diagnostic proposé par Dekker.

IV.1 L'accident du vol 447 d'Air France

En juillet 2012, trois ans après le crash du vol 447 Rio-Paris d'Air France, le rapport du Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile (BEA) est publié [BEA, 2012]. Cette enquête montre qu'à la suite d'une panne des sondes Pitot qui mesurent la vitesse de l'avion, le pilote automatique s'est désactivé et l'avion est entré en décrochage. Les pilotes n'ont alors pas identifié le décrochage de l'avion alors que l'alarme de décrochage a retenti plus de 75 fois. Pour bien comprendre l'erreur des pilotes, il faut savoir que lorsqu'un appareil est en décrochage, la bonne manœuvre à effectuer est de pousser le manche (c'est-à-dire agir sur la commande de vol qui fait basculer l'appareil vers l'avant (figure IV.1)). Au contraire, lorsque l'appareil est en survitesse, il faut tirer sur le manche pour relever le nez de l'appareil [Conversy et al., 2014]. C'est la confusion entre ces deux situations qui a conduit l'appareil à s'écraser.

Le Bureau d'Enquêtes montre que quatre dispositifs principaux ont joué un rôle dans cette confusion :

— Le *directeur de vol* (« Flight Director » ou « FD ») qui indique au pilote quelle manœuvre effectuer pour rejoindre la trajectoire programmée ;

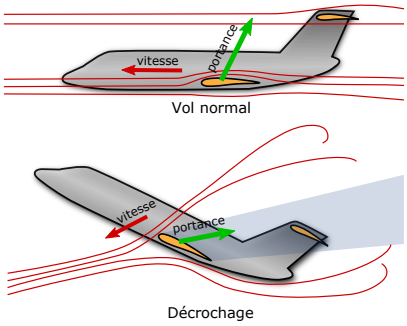


Figure IV.1 – Décrochage (*source : Caliver, Wikimedia Commons*)

— L'*alarme de décrochage* (« stall ») qui se déclenche lorsque l'appareil entre en décrochage (perte de portance entraînant une chute à incidence élevée, voir figure IV.1) ;
— L'altimètre qui donne une *vitesse verticale* (« Vertical Speed »), indicateur de la chute de l'appareil ;
— La sonde Pitot, défectueuse au moment du crash, qui indique la *vitesse* de l'avion (et donc une éventuelle sur-vitesse).

À ces dispositifs s'ajoute un phénomène de vibration appelé « buffet » sur le manche de contrôle de l'appareil qui a aussi joué un rôle dans la confusion entre les deux situations.

Nous résumons la situation de l'accident à partir de l'enquête du BEA et de ces cinq facteurs sur quatre pas de temps :

(t=1) Les indicateurs de vol affichent une accélération soudaine, l'alarme de décrochage s'active et une vibration de type « buffet » est ressentie. Le pilote tire sur le manche.
(t=2) L'alarme de décrochage et la vibration s'arrêtent, la vitesse verticale augmente (perte d'altitude) et le directeur de vol demande de tirer le manche. Le pilote tire le manche.
(t=3) L'alarme de décrochage est toujours désactivée, la vitesse verticale continue d'augmenter et le directeur de vol demande toujours de tirer le manche. Le pilote pousse le manche.
(t=4) L'alarme de décrochage se rallume, la vitesse verticale augmente et le directeur de vol demande toujours de tirer le manche. Le pilote tire le manche.

Pour expliquer les décisions sur chacun de ces pas de temps, le BEA propose les explications suivantes :

1. « le buffet, pouvant être associé dans son esprit à de la haute vitesse » (p. 186)
2. « il est possible qu'un phénomène de sélectivité attentionnelle ait réduit sa capacité de perception de l'alarme [de décrochage]. » (p.188)

3. « De plus, la présence du directeur de vol conduisant à afficher une assiette à cabrer [tirer sur le manche] a pu conforter le [pilote] dans l'idée que l'alarme de décrochage n'était pas pertinente. » (p.187)

Ainsi, pour le BEA, la confusion entre la situation de survitesse et de décrochage est dûe à des informations contradictoires que pouvait observer le pilote ainsi qu'une attention trop grande du pilote à la possibilité d'une situation de survitesse.

Nous allons dans la suite de cette thèse nous reposer sur cet exemple afin d'illustrer notre approche. Dans la section suivante, nous allons aborder les hypothèses que nous faisons dans notre approche pour se rapprocher du diagnostic à la Dekker et définir les différents concepts nécessaires pour effectuer un tel diagnostic.

IV.2 Définition de l'approche de diagnostic

Nous avons pu voir à travers le chapitre II que [Dekker, 2006] propose une méthode pour déterminer les causes d'une prise de décision erronée. De plus, nous avons pu voir section III.4, que nous recherchons en outre à formaliser cette méthode en logique pour proposer une approche formelle du diagnostic de l'erreur humaine. Nous allons voir tout d'abord que formaliser la méthode de Dekker nécessite à la fois de poser des hypothèses et de définir ce que Dekker entend par "point de vue" de l'agent. Nous proposerons ensuite une définition d'un diagnostic en deux temps : un modèle d'explication et un modèle d'évaluation. Enfin, nous définirons le principe général de ces deux modèles.

IV.2.1 Hypothèses à la Dekker

Nous avons vu section II.1 que [Dekker, 2006] décrit le raisonnement d'un enquêteur par le fait de :

(1) se placer du point de vue de l'agent étudié ;
(2) trouver les décisions incohérentes avec le point de vue de l'agent ;
(3) déterminer les croyances rationnelles avec ces décisions.

Notre approche du diagnostic se base sur ce principe et nous en déduisons trois hypothèses importantes dans notre approche. La première découle directement de la première étape (1) :

Hypothèse 1 (H_1) : La modélisation de la situation à analyser doit se faire du point de vue d'un agent

Cette hypothèse nous permet d'éviter de tomber dans les pièges décrits par Dekker (voir section II.3) si nous nous plaçons d'un point de vue omniscient. De plus, nous pensons que la modélisation du point de vue de l'agent dans le cas de l'étude

d'accident d'aviation est atteignable avec l'utilisation des boîtes noires. Nous allons aborder plus précisément cette notion de point de vue dans la section suivante.

La deuxième découle de la notion de décision rationnelle et incohérente de Dekker (*i.e* (2) et (3)) :

Hypothèse 2 (H_2) : Cohérence logique et rationalité sont équivalentes

Nous faisons cette hypothèse afin de considérer tout comme Dekker qu'une décision prise par un agent ne peut pas être en contradiction avec ses croyances. Ainsi, un système logique représentant les croyances et la décision de l'agent doit être impérativement satisfaisable (*i.e* cohérent logiquement). De plus, cela implique qu'une décision incohérente est équivalente à un système logique insatisfaisable qui représente la décision de l'agent et son point de vue.

Enfin, la troisième hypothèse est liée à la manière de déterminer les croyances de l'agent en retrouvant des croyances compatibles avec la décision en fonction du point de vue de l'agent. Nous considérons que :

Hypothèse 3 (H_3) : L'agent ignore les informations incompatibles avec son point de vue et sa décision

Une ignorance est une information dans le point de vue de l'agent que l'agent ne croit pas de façon inconsciente ou consciente. C'est-à-dire que l'agent peut, par exemple, ne pas croire une information du fait qu'il ne l'a pas observée par manque d'attention. Dans ce cas, il n'a pas conscience qu'il a ignoré cette information. Toutefois, dans le cas où l'agent reçoit deux informations contradictoires et choisit d'en préférer une, alors l'agent a conscience qu'il ignore une certaine information pour retrouver la cohérence.

Ces ignorances permettent ainsi de rendre le point de vue compatible avec la décision de l'agent et de déterminer les croyances à la Dekker :

croyances de l'agent = point de vue de l'agent - ignorances.

Il en découle que si le point de vue et la décision sont cohérents, alors les croyances de l'agent et le point de vue de l'agent sont équivalents.

Ces trois hypothèses nous permettent de nous approcher de la vision de Dekker du diagnostic de l'erreur humaine en utilisant un système logique. Nous pouvons voir à travers ces hypothèses que le point de vue de l'agent est central dans l'approche de Dekker. C'est pourquoi nous allons aborder sa définition dans la section suivante.

IV.2.2 Définition du "point de vue" de l'agent

Dekker définit la notion de "point de vue" à un instant t comme l'ensemble des informations disponibles à l'agent ainsi que ses objectifs à cet instant t. L'auteur ne décrit pas précisément ce qu'il entend par "informations disponibles à un instant t".

C'est pourquoi nous considérons que les informations disponibles correspondent à trois principaux concepts :

- Les observations disponibles à un instant t, c'est-à-dire toutes les informations que l'agent peut possiblement voir et entendre à cet instant. Par exemple, l'agent entend une alarme et peut observer que la vitesse augmente.
- Les règles de raisonnement, c'est-à-dire toutes les règles expertes permettant à l'agent d'inférer des croyances et prendre des décisions. Par exemple, l'alarme peut indiquer un décrochage et de ce fait la meilleure décision à prendre est de pousser le manche de contrôle.
- Les croyances précédentes de l'agent, c'est-à-dire ce que croyait l'agent à l'instant $t-1$. Par exemple, l'agent croyait précédemment que l'alarme était défectueuse.
- L'historique des actions de l'agent, c'est-à-dire l'ensemble des actions que l'agent a effectuées du début jusqu'à l'instant t. Par exemple, l'agent avait décidé de tirer le manche jusqu'à l'instant t

Pour représenter la notion "d'objectif" pour l'agent à instant t, nous employons le terme de désir tiré des modèles BDI (voir au III.1.4.c) et qui correspond à un état hypothétique que l'agent souhaite satisfaire. C'est-à-dire qu'une décision est considérée incohérente (au sens de Dekker) si elle entre en conflit avec les désirs possibles de l'agent. Par exemple, l'agent peut avoir le désir de ne pas être en décrochage.

Le point de vue de l'agent est donc complexe et repose sur plusieurs aspects, à la fois sur l'instant présent et le passé. Nous avons fait l'hypothèse (H_1), que le modélisateur doit représenter ce point de vue. Nous pensons que dans un contexte où la situation étudiée est un accident d'aviation, cet objectif est réalisable. En effet, grâce aux boîtes noires installées dans les avions qui captent les sons dans le cockpit (alarme, conversations, etc) ainsi que les données de vol (vitesse, position de l'avion, etc), il est possible d'avoir des données suffisantes pour déterminer les observations disponibles à un instant t. L'historique des actions du pilote est aussi donné par la boîte noire (chaque action dans le cockpit est enregistrée). Les règles de raisonnement de l'agent sont des règles expertes que nous considérons comme connues par le modélisateur : un enquêteur du *Bureau d'Enquêtes et d'Analyses* (BEA) est généralement un pilote confirmé et a une connaissance des règles de raisonnement d'un pilote. Les croyances précédentes de l'agent, quant à elles, doivent être déterminées par notre modèle, tout comme celui-ci doit déterminer les croyances à l'instant t. Enfin, les désirs de l'agent sont laissés à l'appréciation du modélisateur. Ce sont des hypothèses que le modélisateur pose sur les objectifs de l'agent. Il existe néanmoins des désirs qui semblent des hypothèses nécessaires dans la modélisation comme le désir de ne pas mourir, de ne pas se retrouver dans une situation de danger, etc.

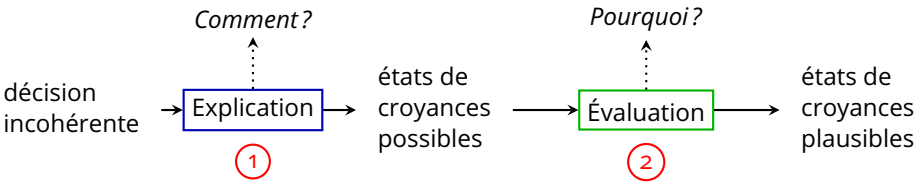


Figure IV.2 – Les deux étapes de l'approche

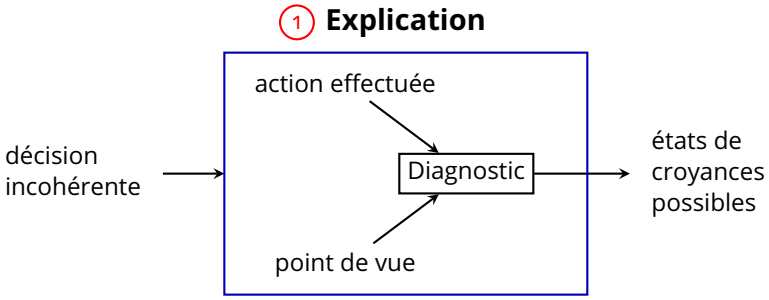


Figure IV.3 – Modèle d'explication

IV.2.3 Définition de l'approche générale

Nous avons conclu section III.4 que le diagnostic de l'erreur humaine devait passer par deux phases importantes. Notre approche du diagnostic va donc être une méthode en deux temps (voir figure IV.2) :

(1) Déterminer les causes d'une prise de décision erronée en passant par un modèle dit d'**explication**. C'est-à-dire un modèle de diagnostic inspiré de la méthode de Dekker permettant d'offrir des explications possibles du comportement de l'agent sur le plan logique.

(2) Comprendre ces causes par rapport à la psychologie humaine (d'un point de vue cognitif) en passant par un modèle dit d'**évaluation**. C'est-à-dire un modèle qui évalue si une explication est plausible ou non en fonction du fait qu'elle correspond à un biais cognitif.

Concrètement, pour un instant t, le modèle d'*explication* prend en entrée l'action et le point de vue de l'agent à t (voir figure IV.3). Les croyances possibles sont alors calculées en effectuant un diagnostic proposé par Dekker. C'est-à-dire à partir du point de vue de l'agent, déterminer les décisions incohérentes et à partir de là construire des croyances rationnelles avec la décision. Nous parlerons de "diagnostic à la Dekker" par la suite.

Ces croyances sont données ensuite au modèle d'*évaluation* (2) qui donne en sortie une mesure de plausibilité sur ces croyances en fonction de la présence ou non de biais cognitifs (voir figure IV.4). La sortie pour un instant t de notre modèle est alors les croyances de l'agent trouvées et la plausibilité de ces croyances.

Nous allons dans les prochaines sections détailler le principe de ces deux modèles.

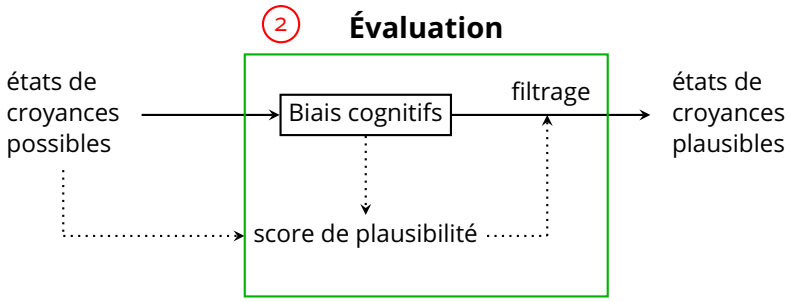


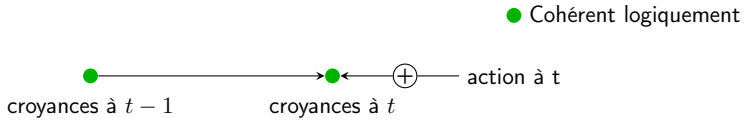
Figure IV.4 – Modèle d'évaluation

IV.2.4 Explication

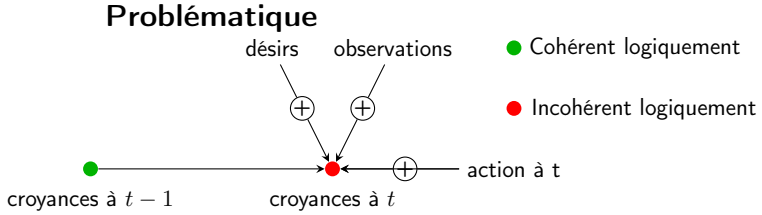
Nous avons posé les différentes hypothèses pour notre modèle et défini les concepts au sein du "point de vue" de l'agent. Nous allons ici décrire le modèle d'*explication* qui repose sur la problématique de la formalisation du diagnostic à la Dekker. L'objectif principal de ce diagnostic est de retrouver les croyances de l'agent qui sont cohérentes avec les décisions de l'agent. Nous allons considérer que chaque action effectuée par l'agent est une décision.

Considérons que nous souhaitons diagnostiquer l'action au pas de temps t. Nous cherchons alors à déterminer des croyances cohérentes logiquement au temps t avec l'action au même pas de temps (et qui sont donc rationnelles au sens de Dekker, *cf* H_2).

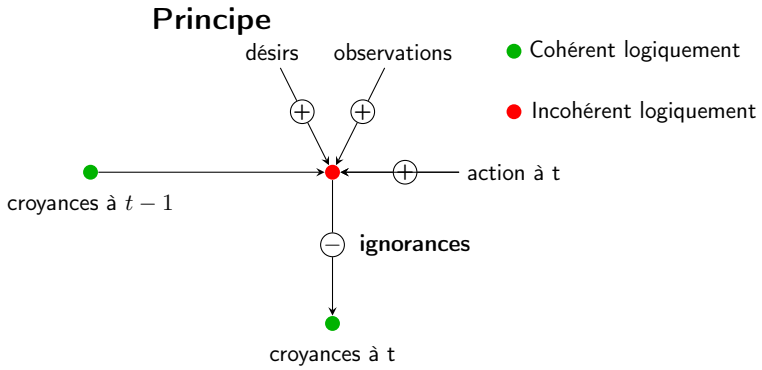
Nous partons donc des croyances précédentes $(t-1)$ de l'agent cohérent logiquement (et donc rationnel au sens de Dekker, *cf* H_2) et cherchons à déterminer des croyances cohérentes logiquement au temps t avec l'action du pas de temps t :



Les croyances de l'agent à un instant t se construisent à partir du point de vue de l'agent (*cf* H_1 et H_3), c'est-à-dire des croyances précédentes, actions précédentes, observations et désirs. Le problème est que cette construction peut être incohérente avec l'action effectuée au temps t (nous avons une décision incohérente au sens de Dekker) :



À partir de là, nous devons retrouver la cohérence dans les croyances du temps t afin d'avoir des croyances rationnelles. L'idée est alors de déterminer les ignorances que l'agent a effectuées pour retrouver la cohérence. Ainsi le point de vue de l'agent auquel on retire les ignorances correspond aux croyances à l'instant t.



Il est possible toutefois que plusieurs ignorances soient possibles pour retrouver la cohérence, **il n'y a donc pas une solution de croyances à un instant t mais plusieurs états de croyances possibles**.

Ce principe de trouver les ignorances de l'agent et par conséquent ses croyances est le cœur de notre approche. Cela permet d'offrir une explication de la décision de l'agent à un temps t. Nous verrons comment nous nous aidons des solutions proposées dans la littérature informatique pour trouver les ignorances dans les chapitres suivants.

Toutefois, ce diagnostic à la Dekker ne suffit pas, car il n'offre pas une compréhension de la raison de ces ignorances (voir la discussion sur le pourquoi du comment section II.2) afin de faire une comparaison entre les états de croyance possibles d'un agent. C'est pourquoi notre approche repose sur un deuxième modèle visant l'objectif d'évaluer la plausibilité d'une explication.

IV.2.5 Évaluation

Par le diagnostic à la Dekker, nous trouvons à la fois les croyances de l'agent et les faits qu'il a possiblement ignorés (que nous appelons "ignorances") par rapport à son point de vue à un instant t. Sachant cela, la question est de comprendre pourquoi ces ignorances ont été effectuées et d'évaluer si l'explication est plausible ou non. Nous savons que l'étude des biais cognitifs que nous avons abordée section II.3 permet d'offrir une explication cognitive à une prise de décision erronée. Nous devons donc, à partir des ignorances et des croyances trouvées à un instant t, déterminer si les ignorances dans ce contexte correspondent à un biais afin de déterminer si l'explication est plausible d'un point de vue cognitif.

Nous avons vu section II.3 qu'il n'existe pas de consensus sur une taxonomie des biais cognitifs et par conséquent de caractéristiques précises pour déterminer les

biais. C'est pourquoi ce deuxième modèle d'*évaluation* de notre approche nécessite dans un premier temps de :

(1) Définir une taxonomie formelle des biais cognitifs

Cette taxonomie doit proposer des caractéristiques sur un système logique qui permettent d'identifier le pourquoi d'une ignorance, en se reposant seulement sur les concepts au sein du modèle formel. Par exemple, une ignorance d'un agent peut être incohérente logiquement avec un des désirs de l'agent, l'ignorance a été déjà effectuée précédemment, etc. Nous verrons plus précisément dans le chapitre VII comment cette taxonomie est construite et les différentes caractéristiques possibles d'une ignorance.

Une ignorance peut être alors identifiée par un ensemble de caractéristiques que nous retrouvons dans la taxonomie formelle. Un biais cognitif pour une ignorance d'une proposition logique φ, est alors défini comme une fonction booléenne $bias(B, \varphi)$, prenant en paramètre les croyances de l'agent B et l'ignorance φ et retourne vraie si une combinaison de n caractéristiques de la taxonomie est satisfaite et correspond au biais cognitif :

$$bias(B, \varphi) \leftrightarrow \begin{cases} \text{caractéristique n°}1 \text{ est vraie} \\ \vdots \\ \text{caractérisique n°}n \text{ est vraie} \end{cases}$$

Ainsi dans un deuxième temps, à partir des définitions des biais de la taxonomie, nous faisons une :

(2) Correspondance entre aucune, une ou plusieurs définitions de biais et une ignorance.

Ce lien entre une définition formelle d'un biais et une ignorance d'une proposition logique nous permet de conclure qu'il existe une explication cognitive à cette ignorance. Par conséquent, les croyances de l'agent résultant de cette ignorance ont plus de chance d'être réaliste.

Enfin, dans un troisième et dernier pas de temps, nous devons :

(3) Évaluer la plausibilité d'un état de croyance en fonction des biais cognitifs trouvés.

Chaque état de croyance possible à un instant t ayant plusieurs ignorances et chaque ignorance ayant plusieurs biais cognitifs possibles, nous devons définir une fonction d'évaluation permettant de comparer les états de croyance en fonction des biais trouvés. Cette fonction doit donc pouvoir offrir un ordre de plausibilité sur l'ensemble des états de croyance possibles retournés par le modèle *d'explication* en charge du diagnostic à la Dekker. Par exemple, si le modèle *d'explication* retourne quatre états de croyance possibles, nous pouvons avoir un ordre de plausibilité comme :

$$B^{(3)} \geq B^{(1)} > B^{(2)} > B^{(4)}$$

Les états de croyance (3) et (1) sont plus plausibles que l'état (2) qui est plus plausible que l'état (4). Nous proposerons une première fonction allant dans ce

sens chapitre VII et soumettrons quelques idées pour améliorer cette fonction.

IV.3 Conclusion

Notre approche repose sur deux modèles :

— Un modèle dit d'*explication* qui détermine les croyances possibles d'un agent à un instant t en utilisant un diagnostic à la Dekker.

— Un modèle dit d'*évaluation* qui utilise les biais cognitifs pour déterminer un ordre de plausibilité sur les mondes possibles qui résultent du modèle *d'explication*.

Nous avons vu que le modèle *d'explication* repose sur la recherche de la cohérence dans un système logique en retrouvant les ignorances possibles de l'agent. Ces ignorances combinées à l'état de croyance de l'agent offrent une explication sur comment l'agent est arrivé à prendre la décision d'une certaine action.

Le modèle *d'évaluation* quant à lui repose sur la définition d'une taxonomie formelle afin de déterminer les caractéristiques sur les ignorances qui identifient un biais. L'identification des biais permet de définir un ordre de plausibilité sur les états de croyance de l'agent et ainsi de déterminer l'état de croyance qui permet d'offrir à la fois le comment et le pourquoi de l'action effectuée par l'agent la plus réaliste.

Dans ce chapitre, nous avons introduit de manière générale l'approche abordée pour le diagnostic d'erreur humaine dans cette thèse. Dans les prochains chapitres, nous verrons comment nous utilisons les outils formels de la littérature vus au chapitre III pour proposer un modèle formel de notre approche.

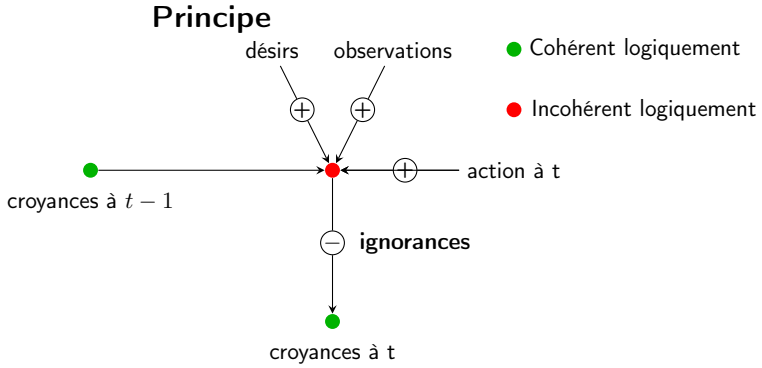
V - Le modèle d'explication

Nous avons vu dans le chapitre précédent que notre approche repose sur un modèle *d'explication* et d'*évaluation*. Le premier se charge de trouver le comment de l'action effectuée par un diagnostic à la Dekker, le deuxième de trouver le pourquoi de l'action par les biais cognitifs.

Dans ce chapitre, nous allons nous attarder sur la formalisation du modèle *d'explication*. Le chapitre est composé de quatre sections. La première définit le langage logique utilisé pour décrire la situation d'accident que le modèle cherche à diagnostiquer. La deuxième section définit les états de croyances de l'agent. La troisième section aborde l'algorithme de diagnostic de ce modèle. Enfin, la quatrième section présente une application de l'algorithme sur le cas d'étude du vol 447, introduit section IV.1.

V.1 Le modèle

Afin de diagnostiquer la décision de l'agent, nous recherchons une formalisation logique du principe vu sous-section IV.2.4 :



Il est donc nécessaire d'avoir un langage logique assez expressif pour représenter tous les concepts présents dans cette problématique : les croyances passées, présentes, les désirs, observations et actions. Nous allons dans un premier temps définir la syntaxe logique utilisée puis dans un deuxième temps à l'aide de cette syntaxe définir les différents concepts utilisés dans le modèle d'*explication*.

V.1.1 Syntaxe

Nous basons notre modélisation sur un langage propositionnel logique où chaque proposition logique est indicée temporellement afin d'exprimer l'instant auquel est considérée la proposition. Pour cela, nous définissons l'ensemble $\mathcal{S}$ des symboles de propositions de notre langage. Nous notons $\mathcal{P}$ l'ensemble des propositions logiques qui correspondent à des symboles de $\mathcal{S}$ indicés temporellement. Pour simplifier l'écriture, tout indice temporel libre est considéré implicitement comme quantifié universellement :

$$p_t \quad \Leftrightarrow \quad \forall t \quad p_t$$

avec $p \in \mathcal{S}$ et $p_t \in \mathcal{P}$. Par exemple :

$\mathrm{alarm}_t \quad \rightarrow$ l'alarme sonne au temps t
$\mathrm{cloud}_{t+1} \quad \rightarrow$ il y a des nuages au temps $t+1$

Nous définissons alors un littéral φ_t par :

$$\varphi_t ::= p_t \mid \neg p_t$$

où $p_t \in \mathcal{P}$. Un littéral est dit négatif quand la proposition du littéral est précédée de la négation logique (*i.e* $\neg$).

Nous définissons le langage $\mathcal{L}_0$ par la grammaire suivante :

$$\alpha ::= \varphi_t \mid \bot \mid \top \mid \neg\alpha \mid \alpha \wedge \alpha \mid \alpha \vee \alpha$$

où α est une formule valide de $\mathcal{L}_0$. La proposition $\bot$ est toujours fausse et la proposition $\top$ est toujours vraie.

Nous définissons aussi un ensemble de symboles $\mathcal{S}^{\mathcal{A}}$ qui représente les actions possibles de l'agent. Les propositions logiques d'action correspondent alors à des symboles d'action dans $\mathcal{S}^{\mathcal{A}}$ et indicés temporellement. Nous notons $\mathcal{A}$ l'ensemble des propositions d'actions :

$$a_t \quad \rightarrow \quad \text{l'action } a \text{ est effectuée à l'instant } t$$

Nous définissons alors le langage $\mathcal{L}$ comme une extension du langage $\mathcal{L}_0$ en ajoutant les actions $\mathcal{A}$, ainsi que les trois opérateurs suivants :

$$\phi ::= \alpha_1 \rightarrow \alpha_2 \mid \{\alpha\} act \mid act \wedge \alpha :: \varphi_t$$

avec $\alpha \in \mathcal{L}_0$, $\alpha_1 \in \mathcal{L}_0$, $\alpha_2 \in \mathcal{L}_0$ et $act \in \mathcal{A}$

- $\alpha_1 \rightarrow \alpha_2$ exprime le fait qu'à partir d'α_1 l'agent infère α_2.
- $\{\alpha\} act$ exprime que α est la précondition de l'action act, c'est-à-dire que pour que l'action act soit effectuée, il est nécessaire que α soit vrai. Par exemple :

$\{\neg \mathrm{locked}_t\}\, \mathrm{doOpen}_t$	l'action d'ouvrir (doOpen_t) nécessite que la porte ne soit pas verrouillée ($\neg \mathrm{locked}_t$).

— $act \wedge \alpha :: \varphi_t$ exprime que φ_t est l'effet de l'action act quand α est vraie, c'est-à-dire que φ_t devient vraie après que l'action act est effectuée sachant α vraie. Par exemple :

$$\mathrm{doOpen}_t \wedge \neg\, \mathrm{blocked}_t :: \mathrm{open}_{t+1}$$

l'action d'ouvrir (doOpen_t) si la porte n'est pas bloquée ($\neg\, \mathrm{blocked}_t$) a pour conséquence que la porte soit ouverte au pas de temps suivant (open_{t+1}).

Ainsi, le langage $\mathcal{L}$ nous permet d'écrire un ensemble de formules et propositions logiques permettant de décrire les concepts utilisés dans notre modèle. Nous allons dans la prochaine section définir chacun de ces concepts grâce à ce langage logique.

V.1.2 Éléments du modèle

Nous avons vu sous-section IV.2.2 les différents concepts nécessaires pour représenter le "point de vue" de l'agent et qui rentrent en jeu dans le principe de notre approche. De plus, nous avons dans la section précédente défini un langage logique $\mathcal{L}$ permettant de décrire des propositions, formules logiques et actions dans le temps. Nous allons dans cette section définir un modèle que nous notons $\mathcal{M}^l = \langle \mathcal{O}bs, \mathcal{R}, \mathcal{I}nit, \mathcal{T}, \mathcal{D} \rangle$ composé des concepts du "point de vue" de l'agent qui sont définis à travers des ensembles de propositions logiques valides dans le langage $\mathcal{L}$. Ce modèle correspond alors à la description formelle du "point de vue" d'un agent sur une situation d'accident. Chacune de ces propositions pourra alors être ignorée pour retrouver la cohérence et déterminer les croyances (voir sous-section IV.2.4) sauf exception que nous préciserons pour les concepts concernés. Nous allons dans cette section définir formellement chaque élément du modèle.

V.1.2.a Les observations

Les observations que nous notons $\mathcal{O}bs$ est une suite d'observation dans le temps : $\mathcal{O}bs = \{Obs_1 \dots Obs_t\}$. Chaque ensemble Obs_k est un ensemble de littéraux positif ou négatif de $\mathcal{L}$ et représente l'ensemble des littéraux possiblement observés par l'agent au pas de temps k. Par exemple :

$\mathrm{alarm}_1 \in Obs_1$	indique que l'agent peut observer au temps 1 que l'alarme sonne au temps 1.
$\neg\, \mathrm{cloud}_2 \in Obs_1$	indique que l'agent peut observer au temps 1 une prévision de la météo lui indiquant qu'il n'y aura pas de nuages au temps 2.

Un littéral qui n'est pas renseigné dans une observation est considéré comme inconnu du point de vue de l'agent : il ne peut pas observer si le littéral est positif ou négatif.

V.1.2.b Les règles de raisonnements

Les règles de raisonnement sont représentées par un ensemble $\mathcal{R}$ de formules de $\mathcal{L}$ avec un indice temporel t libre. Par commodité, chaque élément de $\mathcal{R}$ (donc chaque règle) est identifiée par un littéral positif R^k. Par exemple :

$$R^k \equiv \mathrm{alarm}_t \rightarrow \mathrm{stall}_t$$

La règle de raisonnement R^k correspond à la règle qui indique que lorsque l'alarme sonne, cela indique un décrochage de l'appareil. Nous notons R^k_1, le littéral qui correspond à la règle R^k avec $t = 1$:

$$R^k_1 \equiv \mathrm{alarm}_1 \rightarrow \mathrm{stall}_1$$

Nous distinguons un sous-ensemble de règles dans les règles de raisonnement que nous nommons *règles de cohérence* et qui correspondent à des règles très importantes dans le raisonnement de l'agent.

Les règles de cohérence que nous notons $R^C \subset \mathcal{R}$ sont des règles de raisonnement que l'agent ne peut pas ignorer et qui sont nécessaires pour avoir un raisonnement a minima réaliste. Généralement ce sont des règles de raisonnement qui traduisent un phénomène physique que l'agent ne peut pas outrepasser. Par exemple, la règle R^k :

$$R^k \equiv \mathrm{pull}_t \wedge \mathrm{push}_t \rightarrow \bot$$

qui indique que l'agent ne peut pas pousser (push) et tirer (pull) le manche de contrôle en même temps. Cette règle ne peut en aucun cas être ignorée par l'agent. De ce fait $R^k \in R^C$.

V.1.2.c Les croyances initiales

Les croyances initiales que nous notons $\mathcal{I}nit$ forment un ensemble de littéraux $\varphi \in \mathcal{P}$ positifs ou négatifs qui représentent les croyances au pas de temps 0 de l'agent. Par exemple, si nous avons :

$$\mathcal{I}nit = \{\neg\, \mathrm{alarm}_0, \mathrm{cloud}_1\}$$

alors l'agent croit au temps 0 qu'il n'y a pas d'alarme, mais qu'il y aura des nuages au prochain pas de temps.

Nous considérons qu'un littéral qui, à partir des croyances initiales, observations et des règles de raisonnement, ne peut être déduit est considéré comme inconnu du point de vue de l'agent comme cela est souvent proposé dans les logiques épistémiques non-monotones telles que que la logique autoépistémique [Moore, 1985].

V.1.2.d La trace

La trace que nous notons $\mathcal{T}$ est un ensemble d'action $\mathcal{T} = \{a_1, \ldots, a_t\}$ où chaque $a_k \in \mathcal{A}$ et représente l'action effectuée au pas de temps k par l'agent. Par exemple, la trace suivante :

$$\mathcal{T} \quad = \quad \{\mathrm{pull}_1, \mathrm{pull}_2, \mathrm{push}_3\}$$

indique que l'agent a tiré deux fois avant de pousser le manche de contrôle.

Nous nous limitons à une action par pas de temps afin de garantir que l'explication que nous trouvons à un pas de temps t se limite à expliquer une seule décision de l'agent (une seule action).

V.1.2.e Les désirs

Les désirs que nous notons $\mathcal{D}$ sont un ensemble de littéraux positif ou négatif avec un indice temporel t libre, donc applicables à tout instant t. Nous nous inspirons là encore des logiques BDI et considérons que l'agent cherche à satisfaire ces littéraux. C'est-à-dire que si le littéral φ est un désir alors les croyances de l'agent doivent être cohérentes avec φ ou retrouver la cohérence en ignorant des observations, des croyances, des règles ou même le désir φ. Les désirs dans le cas d'une modélisation d'accident vont être généralement utilisés de manière aversive (un littéral négatif) du fait que le modélisateur peut plus facilement poser des hypothèses sur les états que l'agent souhaite éviter que les états qu'il souhaite atteindre. Par exemple, dans le cas du vol 447, nous pouvons faire l'hypothèse que le pilote avait pour désir :

$$\mathcal{D} \quad = \quad \{\neg\,\mathrm{stall}_t, \neg\,\mathrm{overspeed}_t\}$$

de ne pas être en décrochage ni en survitesse, deux situations très dangereuses pour un pilote d'avion. Toute situation dans laquelle il croit être en décrochage ou en survitesse est incohérente.

V.1.3 Conclusion

Nous avons pu définir un modèle $\mathcal{M}^l = \langle \mathcal{O}bs, \mathcal{R}, \mathcal{I}nit, \mathcal{T}, \mathcal{D} \rangle$ composé des différents concepts du "point de vue" de l'agent, qui sont nécessaires pour déterminer ses croyances en définissant différents ensembles de formules logiques et littéraux du langage $\mathcal{L}$ défini sous-section V.1.1. Nous devons maintenant, à partir de ces ensembles, définir formellement les croyances d'un agent à un instant t à partir du principe d'ignorance que nous avons décrit sous-section IV.2.4. La prochaine section définit formellement un état de croyance et donne une sémantique sur cet état.

V.2 État de croyances

Nous avons vu sous-section IV.2.4 que trouver les croyances d'un agent à un instant t consistait finalement à retrouver la cohérence avec une action au pas de temps t en ignorant des propositions dans le système composé :

— des croyances précédentes ;
— des désirs ;
— des observations.

Nous avons pu définir ces éléments sous-section V.1.2 à partir du langage $\mathcal{L}$ défini sous-section V.1.1. À partir de ces définitions, nous proposons la formalisation d'un état de croyances à un instant t en nous inspirant de l'approche base de données de Shoham (voir au III.1.4.c). En effet, nous allons considérer chaque ensemble décrit sous-section V.1.2 comme une base avec ses propres contraintes. Pour la vérification du respect de ces contraintes et du fait que notre modèle s'appuie sur une logique propositionnelle, nous utilisons un solveur SAT qui permet de vérifier la satisfaisabilité d'un système logique. Nous aborderons plus précisément la sémantique de notre logique dans la section suivante.

— **Contrainte de cohérence** : nous imposons comme contrainte que les ensembles $\mathcal{D}, \mathcal{R}$ et $\mathcal{I}nit$ indépendamment soient cohérents logiquement, c'est-à-dire que les ensembles sont satisfaisables. L'agent n'a donc pas de désirs contradictoires (contrairement au modèle BDI). Les règles de raisonnement ne sont pas contradictoires entre elles et les croyances initiales de l'agent sont consistantes.
— **Contrainte d'unicité** : nous imposons comme Shoham l'unicité d'une action sur un pas de temps t, c'est-à-dire que l'ensemble $\mathcal{T}$ ne contient qu'une seule action avec un indice temporel t. Cette unicité d'action nous permet de garantir que les croyances de l'agent à un temps t expliquent une unique décision de l'agent.

Nous n'imposons aucune contrainte sur l'ensemble des observations Obs_t et considérons que l'agent peut potentiellement observer directement des contradictions dans le monde. Par exemple deux outils indépendants dans le cockpit qui lui indiquent des vitesses contradictoires, possiblement à cause d'une panne de l'un des deux outils de navigation.

À partir de là, l'état de croyances d'un agent au temps t que nous notons B_t correspond à une base composée des désirs $\mathcal{D}$, des observations Obs_t, des croyances précédentes B_{t-1} et d'une action effectuée a_t qui a, elle aussi, une contrainte de cohérence et une contrainte d'unicité. C'est une base cohérente, car nous cherchons un raisonnement rationnel de l'agent (*cf* H_2). De plus, il y a une contrainte d'unicité pour là encore s'assurer que les croyances permettent d'expliquer une unique décision de l'agent. Nous notons $\mathcal{I}$ un ensemble d'ignorances permettant que la contrainte de cohérence soit satisfaite. L'état de croyances B_t

correspond alors à :

$$B_t = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup a_t \cup \mathcal{R}\} \setminus \mathcal{I}$$

où chaque B_t est cohérent.

La définition d'un état de croyances est donc récursive et se définit à partir d'un état de croyances précédent soumis aux mêmes contraintes. Du fait de la récursivité, l'initialisation se fait au temps 0 par la définition suivante :

$$B_0 = \{\mathcal{I}nit \cup \mathcal{D} \cup \mathcal{R}\} \setminus \mathcal{I}$$

Avant de nous intéresser à la méthode pour trouver $\mathcal{I}$, les ignorances permettant de respecter la contrainte de cohérence et de retrouver les croyances de l'agent, nous allons nous attarder sur la sémantique de ces états de croyances et notamment comment nous capturons sémantiquement qu'une proposition est inconnue par l'agent.

V.2.1 Sémantique

Nous avons pu voir dans la section précédente qu'un état de croyances n'est rien d'autre qu'une base respectant une contrainte de cohérence et d'unicité. Nous devons toutefois déterminer ce que l'agent croit ou non dans cette base du fait qu'il peut utiliser les règles de raisonnement $\mathcal{R}$ pour déduire de nouvelles propositions. Nous avons déjà introduit succinctement que nous utilisons un solveur SAT afin de déterminer si les contraintes de la base sont respectées. Nous allons baser notre sémantique sur l'utilisation de ce solveur. Pour cela, nous introduisons la notation $B_t \vdash \varphi$ qui signifie que φ est la conséquence logique du système logique B_t en référence à la logique propositionnelle standard. Par exemple, pour le modus ponens :

$$\frac{B_t \vdash \alpha \rightarrow \varphi, \quad B_t \vdash \alpha}{B_t \vdash \varphi}$$

Il nous faut toutefois définir plus précisément les deux opérateurs d'action que nous avons introduit dans notre langage $\mathcal{L}$ (voir sous-section V.1.1) :

— Pour les préconditions d'action :

$$\frac{a_t \in B_t, \quad \{\varphi\} a_t \in B_t}{B_t \vdash \varphi}$$

Nous supposons que si l'agent a fait l'action a_t, alors c'est qu'il la croyait possible, donc que φ était vraie dans B_t.

— Pour les effets d'action :

$$\frac{a_t \in B_t, \quad B_t \vdash \alpha, \quad a_t \wedge \alpha :: \varphi \in B_t}{B_t \vdash \varphi}$$

nous supposons que si l'agent a fait l'action a_t, il en applique les conséquences : conformément à la règle, φ doit être vraie après a_t lorsque α est vraie.

Ainsi $B_t \vdash \varphi$ si et seulement si le solveur retourne que le système logique $B_t \wedge \neg\varphi$ est insatisfaisable. De ce fait, l'état de croyances B_t est incohérent quand B_t est instatisfaisable, ce qui est équivalent à $B_t \vdash \bot$.

Notre objectif est que nous arrivions à une sémantique telle que $B_t \vdash \varphi$ équivaut à ce que l'agent croit φ au temps t. C'est-à-dire que :

$B_t \vdash \varphi$	$\rightarrow$	L'agent croit φ au temps t
$B_t \vdash \neg\varphi$	$\rightarrow$	L'agent croit $\neg\varphi$ au temps t
$B_t \nvdash \varphi$ et $B_t \nvdash \neg\varphi$	$\rightarrow$	L'agent ne sait pas si φ est vrai ou faux

Toutefois nous allons voir dans la prochaine section que du fait de l'utilisation d'un SAT solveur, nous sommes obligés d'introduire de nouvelles propositions afin de capturer cette sémantique.

V.2.2 Les propositons known

V.2.2.a Objectif

Si nous souhaitons capturer sémantiquement que $B_t \nvdash \varphi$ et $B_t \nvdash \neg\varphi$ équivaut à ce que l'agent ne sait pas la valeur de vérité de φ alors cela implique que le SAT solveur retourne que les deux systèmes $B_t \wedge \varphi$ et $B_t \wedge \neg\varphi$, sont tous les deux satisfaisables. Or avec uniquement la logique utilisée, la construction de l'état de croyances introduite précédemment et l'utilisation d'un solveur SAT, nous pouvons construire un exemple dans lequel le SAT solveur retourne insatisfaisable pour $B_t \wedge \neg\varphi$ alors que la proposition n'est pas connue de l'agent. Considérons les règles de raisonnement et les observations suivantes :

$$\mathcal{R} \equiv \left\{ \begin{array}{lcl} R^k & \equiv & \varphi_t \rightarrow \neg\psi_t \\ R^l & \equiv & \neg\varphi_t \rightarrow \neg\gamma_t \end{array} \right\} \qquad Obs_t \equiv \{\gamma_t, \psi_t\}$$

Dans cet exemple, l'agent observe les propositions γ_t et ψ_t mais rien sur la proposition φ_t. De plus, les règles de raisonnement de l'agent ne lui permettent pas de déduire quoique ce soit sur φ_t. Nous nous limitons ici aux règles de raisonnement et observations pour construire l'état de croyances pour simplifier l'exemple. De ce fait, il est attendu que l'état de croyances qui résulte de $B_t = \mathcal{R} \cup Obs_t$ permet de déterminer que l'agent ne sait pas la valeur de vérité de φ_t. Toutefois, en interrogeant le solveur SAT sur le système $B_t \wedge \neg\varphi_t$, celui-ci va essayer d'attribuer une valeur de vérité à chacune des propositions du système afin de trouver une solution satisfaisable. Or si une valeur vraie ou fausse est attribuée à φ_t le système est de toute façon insatisfaisable :

— φ_t implique $\neg\psi_t$ alors que ψ_t est observé

— $\neg\varphi_t$ implique $\neg\gamma_t$ alors que γ_t est observé

Il y a donc une incohérence dans l'état de croyances de l'agent alors que si φ_t n'a pas de valeur forcée à vrai ou à faux et est considéré comme inconnue, l'agent n'a aucune raison de déduire $\neg\psi_t$ et $\neg\gamma_t$ par les règles R^l ou R^k.

V.2.2.b Principe

Afin d'avoir la sémantique souhaitée, pas d'incohérence quand φ_t est inconnue, nous introduisons les propositions $known$. L'idée est que les règles de raisonnement ne peuvent être appliquées que si les littéraux de la règle sont connus. Concrètement, pour chaque littéral φ_t du langage $\mathcal{L}$ (voir sous-section V.1.1), nous faisons correspondre une proposition $known(\varphi)_t$ telle que :

$known(\varphi)_t$ est vrai si φ_t est connu (*i.e* l'agent croit φ_t ou $\neg\varphi_t$).

avec φ le symbole de proposition (*i.e* $\varphi \in \mathcal{S}$) du littéral φ_t et t l'indice temporel du même littéral. Chaque littéral φ_t qui apparaît dans une règle $R^k \in \mathcal{R}$ est alors transformé tel que :

$$\begin{array}{lll} \varphi_t & \text{devient} & \varphi_t \wedge known(\varphi)_t \\ \neg\varphi_t & \text{devient} & \neg\varphi_t \wedge known(\varphi)_t \end{array}$$

Ainsi l'exemple précédent devient :

$$\begin{array}{rcl} R^k & \equiv & (\varphi_t \wedge known(\varphi)_t) \rightarrow (\neg\psi_t \wedge known(\psi)_t) \\ R^l & \equiv & (\neg\varphi_t \wedge known(\varphi)_t) \rightarrow (\neg\gamma_t \wedge known(\gamma)_t) \end{array}$$

Par conséquent, un littéral est forcé à être $known$ seulement si celui-ci est la conséquence d'une règle avec une prémisse qui est vraie et dont tous les littéraux ont leurs $known$ correspondant à vrai. Dans l'exemple ci-dessus, rien ne force $known(\varphi)_t$ à vrai, le SAT solveur peut donc attribuer la valeur fausse au littéral $known(\varphi)_t$, ainsi les règles R^k et R^l ne peuvent déduire $\neg\psi$ ou $\neg\gamma$ et l'état de croyances B_t est satisfaisable. Nous avons alors la sémantique attendue : le SAT solveur retourne que $B_t \wedge \neg\varphi$ et $B_t \wedge \varphi$ sont tous les deux satisfaisable alors que φ est inconnu. Par conséquent :

$B_t \nvdash \varphi_t$ et $B_t \nvdash \neg\varphi_t$ $\rightarrow$ Indique que l'agent ne sait pas si φ_t est vrai ou faux

V.2.2.c Sémantique

L'utilisation de ces propositions $known$ pour définir les littéraux connus de l'agent nécessite de définir ce qui est initialement connu par l'agent afin de déclencher la déduction des autres littéraux connus par les règles de raisonnement. Pour cela, nous considérons que : les littéraux des croyances initiales $\mathcal{I}nit$, des observations de l'agent Obs_t et l'action effectuée a_t sont connus par l'agent. Par conséquent, ces littéraux ont directement leur proposition $known$ forcée à vrai. Si nous reprenons l'exemple précédent, nous avons :

$$\mathcal{R} \equiv \left\{ \begin{array}{rcl} R^k & \equiv & (\varphi_t \wedge known(\varphi)_t) \rightarrow (\neg\psi_t \wedge known(\psi)_t) \\ R^l & \equiv & (\neg\varphi_t \wedge known(\varphi)_t) \rightarrow (\neg\gamma_t \wedge known(\gamma)_t) \end{array} \right\}$$

$$Obs_t \equiv \{(\gamma_t \wedge known(\gamma)_t), (\psi_t \wedge known(\psi)_t)\}$$

De ce fait, si un littéral est cru par un agent à un instant t (*i.e* $B_t \vdash \varphi_t$) cela veut dire que :

— soit $\varphi_t \in Obs_t$ ou $\varphi_t \in \mathcal{I}nit$;

— soit φ_t est déduit par les règles $\mathcal{R}$ à partir des observations, croyances initiales ou des actions.

Pour une facilité de lecture des différents exemples de modélisation, nous ne spécifierons pas les propositions $known$ sur les actions, croyances initiales, observations et dans les règles de raisonnement, bien que le lecteur doive les considérer comme présents.

Il est à noter que notre proposition syntaxique et sémantique de la prise en compte de la connaissance de l'agent se rapproche des récents travaux de [Cooper et al., 2021]. Les auteurs considèrent aussi que la connaissance vient de l'observation. Ils proposent alors une logique épistémique légère basée sur un opérateur $S_i\varphi$ qui indique que l'agent i observe la valeur de vérité de φ. L'opérateur épistémique $K_i\varphi$ équivaut alors à $S_i\varphi \wedge \varphi$, c'est-à-dire que l'agent i connaît φ, s'il observe la valeur de vérité de φ et que φ est vrai.

V.2.3 Conclusion

Nous avons pu dans cette section définir une formalisation de l'état de croyances d'un agent qui n'est rien d'autre qu'une base de propositions logiques ayant une contrainte de cohérence et d'unicité sur les actions. De plus, nous avons défini la sémantique de l'état de croyances de l'agent en s'appuyant sur des propositions $known$ qui permettent de différencier un littéral connu ou inconnu par l'agent et d'un SAT solver qui nous permet d'interroger l'état de croyances d'un agent pour déterminer ses croyances.

Toutefois, la contrainte de cohérence imposée sur l'état de croyances implique de retirer certaines propositions logiques pour retrouver la cohérence que nous avons définie comme des ignorances sous-section IV.2.4. Ces ignorances permettent de retrouver les croyances de l'agent et ainsi offrir un diagnostic de l'action effectuée par l'agent. Nous formaliserons la recherche de ces ignorances dans la prochaine section.

V.3 Principe du diagnostic

Nous avons vu que la formalisation du diagnostic à la Dekker, qui consiste à retrouver les croyances de l'agent cohérentes avec l'action effectuée par celui-ci en fonction des informations disponibles de son "point de vue", équivaut à partir d'un modèle $\mathcal{M}^l$ du "point de vue de l'agent" de retrouver des ignorances $\mathcal{I}$. En effet, les croyances d'un agent à un instant t correspondent à l'ensemble B_t tel que :

$$B_t = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R}\} \setminus \mathcal{I} \text{ est cohérent}$$

où $\mathcal{I}$ est un ensemble de propositions retirées de l'ensemble afin de retrouver la cohérence. L'objectif principal pour retrouver les croyances et avoir un diagnostic à la Dekker est donc de déterminer $\mathcal{I}$. Toutefois, $\mathcal{I}$ doit trouver une solution à deux types d'incohérences : des incohérences liées à l'action effectuée et des incohérences liées aux observations. Nous allons dans la prochaine section développer ce que représentent ces deux types d'incohérences et faire le lien avec les solutions proposées dans la littérature abordée chapitre III. Nous verrons ensuite comment calculer un état de croyances cohérent à l'aide de ces outils. Enfin, nous présenterons la notion de scénario qui définit une suite d'états de croyances possibles dans le temps.

V.3.1 Deux types d'incohérences, des ignorances différentes mais un même opérateur

Lorsque nous cherchons des ignorances $\mathcal{I}$ afin que B_t soit cohérent, c'est finalement pour corriger deux types d'incohérences :

- Des incohérences liées à l'action effectuée, c'est-à-dire que sans prendre en compte l'action dans B_t celui-ci serait cohérent. Par exemple, l'agent observe l'alarme de décrochage et décide de tirer le manche au lieu de le pousser.
- Des incohérences liées aux observations, c'est-à-dire que l'agent observe des informations contradictoires entre elles par raisonnement. Par exemple, l'agent observe à la fois l'alarme de décrochage et une accélération qui lui indique une survitesse en sachant que les deux situations sont contradictoires.

On doit bien sûr potentiellement corriger une combinaison de ces deux types d'incohérences. C'est-à-dire une situation dans laquelle une action incohérente a été effectuée et que l'agent observe aussi des informations contradictoires.

Nous allons dans les prochaines sections faire le lien entre ces deux problématiques d'incohérences et les solutions de la littérature.

V.3.1.a Incohérence avec l'action

Faisons l'hypothèse ici qu'il n'existe pas d'incohérence liée aux observations et par conséquent que le modèle M doit uniquement corriger des incohérences liées à l'action effectuée. Dans un tel cas, les ignorances que nous recherchons correspondent à une erreur de l'agent qui n'a pas observé une information, n'a pas appliqué une règle de raisonnement, etc. Nous recherchons finalement un diagnostic de l'action effectuée en retrouvant la cohérence dans B_t

Or, ce principe d'un diagnostic par la recherche de cohérence dans un système a été déjà abordé sous-section III.2.3 avec le *consistency-based-diagnosis* de [Reiter, 1987]. Le système à diagnostiquer est composé de trois ensembles :

SD la description du système, ASS les hypothèses des composants fonctionnant normalement, OBS les observations sur le comportement du système. Le diagnostic Δ consiste alors à retirer le minimum d'hypothèses dans ASS pour retrouver la cohérence dans le système :

$$SD \cup OBS \cup (ASS \setminus \Delta) \text{ est cohérent}$$

Nous pouvons donc alors faire une équivalence entre le diagnostic de l'action de notre problématique et un *consistency-based diagnosis*. Dans notre cas :

- la description du système SD correspond aux éléments du modèle $\mathcal{M}^l$, c'est-à-dire les croyances précédentes B_{t-1}, les désirs $\mathcal{D}$, les observations Obs_t et règles de raisonnement $\mathcal{R}$;
- les observations sur le système OBS correspondent à l'action a_t de l'agent ;
- les hypothèses sur le système ASS correspondent non pas à des composants qui fonctionnent normalement, mais des hypothèses sur le fait que l'agent est omniscient sur les observations, applique toutes les règles de raisonnement, satisfait tous ses désirs, etc ;
- le diagnostic Δ correspond aux ignorances $\mathcal{I}$ que nous recherchons.

Ainsi, si nous voulons trouver les ignorances de $\mathcal{I}$ dans le cas d'une incohérence liée à l'action, il suffit d'appliquer un *consistency-based diagnosis*.

V.3.1.b Incohérence liée aux observations

Faisons l'hypothèse ici qu'il n'existe pas d'incohérence liée à l'action effectuée et par conséquent que le modèle M doit uniquement corriger des incohérences liées aux observations. Dans un tel cas, les ignorances que nous recherchons correspondent à des préférences sur les croyances de l'agent. En effet, nous tombons ici sur un problème de *révision de croyance* abordé au III.1.5.a : l'agent fait face à une contradiction et doit abandonner un ensemble de croyances pour retrouver la cohérence. En d'autres termes, l'agent choisit de préférer une croyance plutôt qu'une autre pour garder la cohérence.

Pour résoudre cette incohérence, nous pouvons donc utiliser un opérateur AGM qui permet de calculer la *révision minimale*. Les ignorances correspondront alors aux croyances que l'agent a décidé d'abandonner pour résoudre la contradiction.

V.3.1.c Un même opérateur

Nous avons vu que les incohérences liées à l'action effectuée peuvent être résolues par un opérateur de *consistency-based diagnosis* et les incohérences liées aux observations par un opérateur de révision AGM. Or, nous avons également vu sous-section III.2.4 que ces deux opérateurs étaient équivalents. Par conséquent, les ignorances $\mathcal{I}$ peuvent être trouvées par un même opérateur. De plus, s'il y a la fois des incohérences liées à l'action et aux observations, cet opérateur trouvera un ensemble d'ignorances permettant de corriger les deux types d'incohérences,

bien que les ignorances ne représentent pas la même chose (une erreur ou une préférence sur une croyance). Nous verrons au chapitre VII comment le modèle d'*évaluation* permet de séparer ces deux types d'incohérences et nous allons définir dans les prochaines sections notre opérateur de diagnostic qui permet de les résoudre conjointement.

V.3.2 Définition de l'opérateur de diagnostic

Pour définir notre opérateur de diagnostic, nous avons choisi de nous reposer sur le calcul de *Minimal Correction set* (MCS) qui sont les ensembles minimaux permettant de corriger un système Φ insatisfaisable pour le rendre satisfaisable. Formellement :

> Pour un système $\Phi = \{\phi_1, \phi_2 \dots \phi_n\}$, $M \subseteq \Phi$ est un MCS de Φ si et seulement si :
> — $\Phi \setminus M$ est satisfaisable
> — $\forall \phi_i \in M, (\Phi \setminus M) \cup \{\phi_i\}$ est insatisfaisable

Nous avons fait le choix de nous reposer sur ces ensembles du fait de la proximité de la définition formelle des MCS à la définition du *consistency-based diagnosis*, c'est-à-dire trouver un minimum de propositions à retirer pour retrouver la cohérence. De plus, la littérature compte de nombreux algorithmes pour calculer les MCS [Marques-Silva et al., 2013, Narodytska et al., 2018], dont l'algorithme de [Liffiton et al., 2008] facilement implémentable sur un *Satisfiability modulo theories* (SMT) solveur. Un SMT solveur est simplement une extension des solveurs SAT permettant de raisonner sur des logiques de premier ordre. Tout problème SAT peut être donc implémenté sur un solveur SMT. Nous allons donc définir formellement notre opérateur de diagnostic grâce à l'algorithme de Liffiton et al.

V.3.2.a Algorithme de Liffiton pour le calcul des MCSes

Nous utilisons l'algorithme de Liffiton pour calculer l'ensemble des MCS d'un système Φ. Cet algorithme peut être abstrait en une fonction prenant deux paramètres :

$$\mathfrak{M}(\Phi, screened)$$

où la fonction $\mathfrak{M}$ retourne l'ensemble des MCS du système Φ, Φ étant le système à corriger et $screened \subset \Phi$ étant un sous-ensemble immunisé à la correction. C'est-à-dire que toute proposition dans $screened$ ne peut pas être retournée par $\mathfrak{M}$ dans la même idée que la *screened revision* abordée au III.1.5.a.

Le calcul des MCSes, décrit en détail dans l'Algorithme 1, repose sur le principe suivant :

1. L'ajout de variables y_p dites de sélection sur toutes les propositions $p \in \Phi$ qui ne sont pas dans le sous-ensemble $screened$. Une variable y_p est définie

de telle sorte que sa négation rend la proposition p fausse :

$$\forall p \in \Phi, p \notin screened \quad \neg y_p \rightarrow \neg p$$

L'ajout de ces variables permet de pouvoir ignorer une proposition p en mettant y_p à faux. De ce fait, s'il existe une correction possible pour le système, elle correspond à un ensemble de y_p à faux. Le système est donc SAT tant qu'il y a une correction possible, car il existe une ou plusieurs variables de sélection à faux qui rendent le système satisfaisable.

2. L'algorithme recherche les MCS de manière croissante sur la taille k de la correction possible en commençant par $k = 1$. Pour cela, une contrainte *AtMost* est rajouté à Φ qui consiste à imposer que la somme des variables de sélection à faux soit égale à k. Par exemple, si $k = 2$ alors, nous imposons que la somme des variables de sélection y_p à faux doit être égale à deux.
3. S'il existe une solution satisfaisable avec la contrainte de taille k, alors il existe un MCS de taille k qui correspond à un ensemble de variables de sélection à faux.
4. Afin de veiller à l'aspect minimal d'un MCS, et pour qu'un MCS n'en chevauche pas un autre, nous ajoutons des clauses bloquantes à Φ, ce qui consiste à faire la disjonction des variables de sélection du MCS trouvé. Ainsi, nous excluons le MCS et ses sur-ensembles dans les prochaines solutions, car étant des sur-ensembles d'une correction plus petite, elles ne sont pas minimales. Par exemple, si y_p et y_t constituent un MCS de taille 2, alors nous ajoutons à Φ :

   $$y_p \vee y_t$$

   Ainsi, nous bloquons forcément une proposition appartenant à un MCS à vrai ce qui oblige à trouver un autre MCS. Par exemple, avec la clause bloquante de taille 2 ci-dessus, il est impossible de trouver plus tard une correction de taille 3 qui est égale à : $\neg y_p, \neg y_t, \neg y_i$ car il est nécessaire qu'au moins y_p ou y_t soit vrai du fait de la clause bloquante.
5. Tant que Φ est satisfaisable **avec** la contrainte *atMost* pour la taille k, alors il existe un autre MCS de taille k et nous le recherchons via l'étape 3.
6. Si Φ est satisfaisable **sans** la contrainte *atMost* et avec les clauses bloquantes, cela veut dire qu'il existe toujours une correction minimale de taille k supérieure. Nous la recherchons alors avec l'étape 3 et une taille $k + 1$.
7. Enfin si Φ est insatisfaisable **sans** la contrainte *atMost* alors cela veut dire que toutes les corrections minimales ont été trouvées.

Considérons l'exemple suivant afin d'illustrer l'algorithme avec pour ensemble Φ et $screened$:

$$\Phi \quad = \quad \{\varphi, \psi, \gamma, R^k \equiv \varphi \vee \psi \rightarrow \neg\gamma\} \qquad\qquad screened \quad = \quad \{\gamma\}$$

Algorithm 1 Retourne l'ensemble des MCSes de Φ

$\Phi' \leftarrow$ AddYVars(Φ, $screened$) ▷ Ajout des variables de sélection
MCSes $\leftarrow \emptyset$ ▷ Liste des MCeS
k $\leftarrow$ 1 ▷ Taille du MCS calculé
while SAT(Φ') **do** ▷ Tant qu'il existe des MCeS
  $\Phi'_k \leftarrow \Phi' \wedge$ AtMost($\{\neg y_1, \neg y_2, ..., \neg y_n\}$,k) ▷ Ajout de la contrainte d'ignorance de taille k
  while (newMCS $\leftarrow$ IncrementalSAT(Φ'_k)) **do** ▷ Si un MCS de taille k est trouvé
    MCes $\leftarrow$ MCSes $\cup$ {newMCs} ▷ Ajout dans la liste
    $\Phi'_k \leftarrow \Phi'_k \wedge$ BlockingClause(newMCS) ▷ Ecriture de la clause bloquante du MCS
    $\Phi' \leftarrow \Phi' \wedge$ BlockingClause(newMCS)
  k $\leftarrow$ k+1 ▷ On passe à la taille suivante
return MCSes

Nous ajoutons tout d'abord les variables de sélection :

$$\Phi' = \left\{ \begin{array}{c} \varphi, \psi, \gamma, R^k, \\ \neg y_\varphi \rightarrow \neg\varphi, \neg y_\psi \rightarrow \neg\psi \\ \neg y_{R^k} \rightarrow \neg R^k \end{array} \right\}$$

Nous ajoutons la contrainte de solution de taille 1 pour les variables de sélection à faux :

$$\Phi'_1 = \left\{ \begin{array}{c} \varphi, \psi, \gamma, R^k, \\ \neg y_\varphi \rightarrow \neg\varphi, \neg y_\psi \rightarrow \neg\psi \\ \neg y_{R^k} \rightarrow \neg R^k \\ atMost(\{\neg y_\varphi, \neg y_\psi, \neg y_{R^k}\}, 1) \end{array} \right\}$$

Nous trouvons alors que Φ'_k est satisfaisable et trouvons comme solution :

$\{\neg y_{R^k}\}$ est vrai

Nous ajoutons alors la clause bloquante correspondante sur les deux systèmes :

$$\begin{array}{rcl} \Phi'_1 & = & \Phi'_1 \cup \{y_{R^k}\} \\ \Phi' & = & \Phi' \cup \{y_{R^k}\} \end{array}$$

Φ'_1 est alors insatisfaisable, tous les MCS de taille 1 ont été trouvés. Toutefois, Φ' est toujours satisfaisable, nous devons alors passer à la taille 2 :

$$\Phi'_2 = \Phi' \cup atMost(\{\neg y_\varphi, \neg y_\psi, \neg y_{R^k}\}, 2)$$

On trouve alors comme solution :

$\{\neg y_\varphi, \neg y_\psi\}$ sont vrais

Nous ajoutons alors la clause bloquante correspondante sur les deux systèmes :

$$\begin{array}{rcl} \Phi'_2 & = & \Phi'_2 \cup \{y_\varphi \vee y_\psi\} \\ \Phi' & = & \Phi' \cup \{y_\varphi \vee y_\psi\} \end{array}$$

Φ'_2 et Φ' sont alors tous les deux insatisfaisables ce qui veut dire que tous les MCSes ont été trouvés :

$$MCSes \quad = \quad \{\{R^k\}, \{\varphi, \psi\}\}$$

V.3.3 Définition du diagnostic des croyances

Nous avons pu voir section V.2 que les croyances d'un agent à un instant t correspondaient à l'ensemble :

$$B_t = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R}\} \setminus \mathcal{I}$$

où $\mathcal{I}$ est un ensemble d'ignorances à trouver pour que B_t soit cohérent. De plus, nous avons pu voir section V.3 que ces ignorances pouvaient être calculées grâce à un opérateur à la *consistency-based diagnosis* : $\mathfrak{M}\big(\Phi, screened\big)$ qui retourne l'ensemble des MCS qui corrigent le système Φ sans corriger le sous-ensemble $screened$. Par conséquent, les ignorances $\mathcal{I}$ de l'état de croyances B_t doivent être un résultat de cet opérateur : $\mathcal{I} \in \mathfrak{M}\big(\Phi, screened\big)$. La question est quels sont les paramètres Φ et $screened$ dans notre cadre de correction du système B_t ? La question sous-jacente est quelles sont les propositions que nous considérons que l'agent ne peut pas ignorer ?

En effet, dans notre cas, le système Φ à corriger est l'ensemble des éléments de l'état de croyances :

$$\Phi = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R}\}$$

Pour ce qui est de l'ensemble $screened$ qui ne peut pas être corrigé et donc ignoré par l'agent, nous considérons plusieurs éléments. Tout d'abord, nous avons vu sous-section V.1.2 que le sous-ensemble R^C de $\mathcal{R}$ représentait des règles inviolables par l'agent, généralement sur la physique du monde. De ce fait, nous considérons que ces règles ne peuvent pas être ignorées. Enfin, comme nous recherchons des croyances cohérentes avec l'action de l'agent, nous considérons que l'agent ne peut pas ignorer l'action qu'il a effectuée et croire par exemple qu'il a fait une autre action. Nous considérons que ce type de scénario n'est pas réaliste. De ce fait :

$$screened = \{a_t, R^C\}$$

Par conséquent, la formalisation du diagnostic des croyances d'un agent à un instant t correspond à l'ensemble B_t suivant :

$$\begin{array}{ll} & B_t = \Phi \setminus \mathcal{I} \\ \text{avec} & \mathcal{I} \in \mathfrak{M}\big(\Phi, screened\big) \\ & \Phi = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R}\} \\ & screened = \{a_t, R^C\} \end{array} \tag{V.1}$$

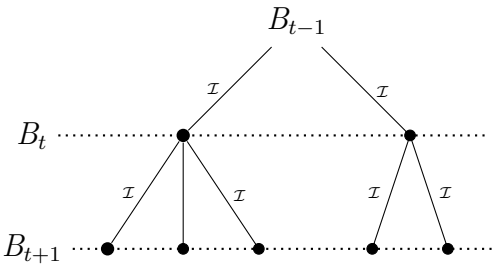


Figure V.1 – Structure en arbre du diagnostic

V.3.4 Définition des scénarios

Par l'équation V.1 nous avons défini le calcul d'un état de croyances pour un pas de temps t. Or, pour un diagnostic complet, il est nécessaire de trouver les états de croyances pour les autres pas de temps du modèle $\mathcal{M}^l$ afin d'avoir un cheminement complet des croyances de l'agent au cours du temps qui explique chaque action dans $\mathcal{T}$ du modèle $\mathcal{M}^l$.

Nous savons d'après l'équation V.1 et la définition de l'opérateur de diagnostic vu sous-section V.3.2 que :

1. la définition de B_t est récursive car définie à partir de l'état précédent B_{t-1} ;
2. il existe plusieurs B_t possibles du fait qu'il existe plusieurs $\mathcal{I}$ retournés par l'opérateur $\mathfrak{M}$.

De ce fait, il en découle que nous avons une structure en arbre : un état de croyances B_{t-1} est le parent de plusieurs états de croyances B_t. Ainsi un nœud dans l'arbre correspond un état de croyances et un niveau de l'arbre correspond à tous les états de croyances possibles à un instant t sachant l'état de croyances précédent B_{t-1} (voir figure V.1). Un diagnostic complet possible pour $\mathcal{M}^l$ est alors un chemin de la racine à une feuille dans cet arbre. Nous nommons ce chemin un *scénario* et le notons S_i. Il est à noter que cette structure en arbre des états de croyances peut être retrouvée si nous ne gardons en mémoire que les ignorances. En effet, un état de croyances B_t par sa définition :

$$B_t = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R}\} \setminus \mathcal{I}$$

n'est rien d'autre que l'union des règles, des croyances initiales et désirs ainsi que toutes les observations et actions du temps 1 au temps t auquel est retiré l'ensemble des ignorances effectuées du temps 0 au temps t. Un état de croyances peut être donc défini par une suite d'ignorances. Nous retrouvons alors la structure en arbre où chaque nœud est un ensemble d'ignorances $\mathcal{I}$ auquel nous pouvons faire correspondre un état de croyances. Les ignorances étant généralement des ensembles plus petits que les états de croyances, ne garder en mémoire que ceux-ci nous permet d'être moins complexe en espace.

Nous notons un nœud $\eta = \langle \rho, \mathcal{I}, t, \mathfrak{N} \rangle$ comme une structure composée d'un ensemble d'ignorances $\mathcal{I}$, d'un pas de temps t, d'un nœud parent ρ et d'un ensemble de nœuds suivants $\mathfrak{N} = \{\eta_1, \ldots, \eta_n\}$. Nous notons r le nœud racine tel que $r = \langle \emptyset, \emptyset, \mathfrak{N} \rangle$. Nous notons $\mathfrak{T}$ l'arbre de diagnostic contenant l'ensemble des nœuds calculés par l'algorithme de diagnostic $\mathfrak{T} = \{r, \eta_1, \ldots, \eta_n\}$. Pour faciliter la lecture, nous notons $\mathcal{I}_t^\eta$ les ignorances du nœud η effectuées au temps t. Nous notons alors la fonction $\mathcal{N}(\mathcal{I}_t^\eta)$ qui retourne l'ensemble des ensembles $\mathcal{I}_t^x$ d'ignorances voisins de $\mathcal{I}_t^\eta$ dans $\mathfrak{T}$. Enfin, nous notons le scénario $S(\mathcal{I}_t^\eta) = (\mathcal{I}_0^a, \mathcal{I}_1^b, \ldots, \mathcal{I}_t^\eta)$ qui est la suite des ensembles d'ignorances dans le chemin du nœud racine au nœud contenant $\mathcal{I}_t^\eta$.

Si nous notons $B_t^{\mathcal{I}^\eta}$, l'état de croyances résultant de l'ignorance $\mathcal{I}_t^\eta$, celui-ci est défini par :

$$B_t^{\mathcal{I}^\eta} = \left\{ \bigcup_{j=1}^{t} \{\mathcal{I}nit \cup \mathcal{R} \cup \mathcal{D} \cup Obs_j \cup \{a_j\}\} \right\} \setminus \left\{ \bigcup_{\mathcal{I} \in S(\mathcal{I}_t^\eta)} \mathcal{I} \right\} \tag{V.2}$$

V.3.5 Conclusion

Dans cette section, nous avons pu proposer une définition formelle du diagnostic des croyances d'un agent à un instant t à partir d'un modèle $\mathcal{M}^l$ du "point de vue" de l'agent. Cette définition se base sur un diagnostic à la *consistency-based diagnosis* en utilisant un opérateur $\mathfrak{M}$ retournant un ensemble de corrections minimal appelé MCS qui correspond dans notre cas à des ignorances de la part de l'agent. Cet opérateur prend en paramètre le système à corriger, dans notre cas le système logique composé de tous les éléments de l'état de croyances, et un paramètre $screened$ qui immunise un sous-ensemble à la correction. Nous considérons qu'un agent ne peut ignorer certaines règles physiques du monde ainsi que l'action qu'il a effectuée. La définition du diagnostic pour un instant t est résumée par l'équation V.1.

Cet opérateur de diagnostic $\mathfrak{M}$ utilise l'algorithme de Lifitton présent dans la littérature que nous avons implémenté à l'aide d'un SMT solveur, plus précisément le SMT solveur de Microsoft : Z3 [Moura et al., 2008]. Notre implémentation de cet algorithme et du modèle de diagnostic en général peut être trouvée sur GitHub [1].

Enfin, un diagnostic complet des croyances d'un agent à partir d'un modèle $\mathcal{M}^l$ correspond à ce que nous nommons des *scénarios* qui sont des suites d'ensembles d'ignorances à partir desquelles les états de croyances peuvent être reconstruits récursivement.

Nous allons section V.5 présenter une application à l'exemple de l'accident du vol 447 abordé section IV.1 pour illustrer le calcul du diagnostic avec cet opérateur. Toutefois, avant toute chose, nous allons dans la prochaine section faire le lien

1. https://github.com/valentinFouillard/SherlockFramework

formel entre notre opérateur de diagnostic $\mathfrak{M}$ et le *consistency-based diagnosis* en utilisant les résultats de la littérature sur le lien entre les opérateurs AGM et ce type de diagnostic que nous avons abordé sous-section III.2.4.

V.4 Équivalence avec un opérateur AGM

Nous avons dans les sections précédentes justifié notre opérateur de diagnostic $\mathfrak{M}$ basé sur le calcul de *Minimal Correction Set* (MCS) par la proximité de la définition d'un MCS et d'un *consistency-based diagnosis*. De plus, du fait, de la relation entre un *consistency-based diagnosis* et une révision de croyance AGM (voir sous-section III.2.4), nous en avons conclu que $\mathfrak{M}$ est à la fois un opérateur de révision AGM et un opérateur de diagnostic. Afin de confirmer cette intuition, nous allons dans cette section montrer l'équivalence entre notre opérateur $\mathfrak{M}$ et une révision de croyance à la AGM.

Nous avons pu voir en sous-section III.1.5 que la révision de croyance minimale d'AGM (notée $*$) pouvait être définie à partir de la contraction minimale d'AGM (notée $\div$) et l'expansion (notée $+$) par la *Levi identity* :

$$K * \varphi = (K \div \neg\varphi) + \varphi$$

Il nous suffit donc de montrer que notre opérateur de diagnostic $\mathfrak{M}$ est équivalent à une contraction AGM suivi d'une expansion. Pour cela, nous allons nous intéresser à la notion de *meet contraction* qui est à la base de la définition de la contraction minimale d'AGM. Nous allons commencer par présenter notre formalisation d'AGM sous l'outil Isabelle, puis nous montrerons que l'opérateur $\mathfrak{M}$ respecte l'axiomatique des opérateurs AGM.

V.4.1 Meet contraction

Les axiomes d'AGM ont été définis à partir d'une approche constructive. Les auteurs d'AGM ont construit des opérateurs dits de *meet contraction* qui permettent d'effectuer une contraction minimale. Dans un second temps, ils ont formalisé ces travaux en définissant une axiomatique pour ces opérateurs. Ainsi, tout nouvel opérateur de *meet contraction* doit respecter à minima les axiomes d'AGM.

Nous allons donc avant toute chose définir ces différentes notions.

Remainders La contraction minimale d'AGM consiste à retirer une proposition φ d'un ensemble de propositions K en s'assurant que l'ensemble retenu de clauses est maximal et que rien de ce qui reste de K ne permet de retrouver φ. En d'autres termes, la contraction cherche un sous ensemble maximal de K qui ne permet pas de déduire φ sans retirer inutilement des clauses. On parle alors de contraction de K par φ.

Pour calculer cet ensemble, les auteurs d'AGM se basent sur la notion de *remainders* qui est l'ensemble des sous-ensembles maximaux de K qui n'impliquent pas φ et est noté : $K \overset{R}{\perp} \varphi$. Formellement :

$$K \overset{R}{\perp} \varphi \equiv \left\{ B \;\middle|\; \begin{array}{l} B \subseteq K \\ B \nvdash \varphi \\ \forall B' \subseteq K, B \subset B' \rightarrow B' \vdash \varphi \end{array} \right\}$$

Fonction de sélection La solution intuitive est de dire que chacun des *remainders* est une solution à la contraction minimale. Cependant, les auteurs de AGM ont montré qu'une telle solution retourne des ensembles « trop grands » en pratique. Pour palier ce problème, ils proposent d'utiliser une fonction de *sélection* $\gamma(K \overset{R}{\perp} \varphi)$ sur les *remainders* qui doit respecter certains axiomes :

(a) Si $K \overset{R}{\perp} \varphi \neq \emptyset$ alors $\gamma(K \overset{R}{\perp} \varphi) \subseteq K \overset{R}{\perp} \varphi$: s'il existe des ensembles qui n'impliquent pas φ alors la fonction de *sélection* γ retourne un sous ensemble de ces ensembles.

(b) Si $K \overset{R}{\perp} \varphi = \emptyset$ alors $\gamma(K \overset{R}{\perp} \varphi) = K$: s'il n'existe pas d'ensemble qui n'implique pas φ alors la fonction de *sélection* retourne K.

(c) $\gamma(K \overset{R}{\perp} \varphi) \neq \emptyset$: la fonction de sélection ne doit pas être vide.

(d) Si $\varphi \longleftrightarrow \psi$ alors $\gamma(K \overset{R}{\perp} \varphi) = \gamma(K \overset{R}{\perp} \psi)$: deux propositions équivalentes doivent mener à la même sélection.

Meet contraction Une *meet contraction* est une construction de la contraction $(\div)$ à partir des *remainders* et d'une fonction de *sélection* γ. Nous notons :

$$K \div^{\gamma} \varphi = \bigcap \gamma(K \overset{R}{\perp} \varphi)$$

tel que $\div^{\gamma}$ est une *meet contraction* qui est définie par une fonction de *sélection* γ. Une *meet contraction* est donc l'intersection des éléments sélectionnés par la fonction de sélection.

La fonction de *sélection* décide des propriétés de la fonction de contraction correspondante. Si la fonction de *sélection* retourne un sous ensemble des *remainders*, nous parlons de *partial meet contraction*. Si la fonction de *sélection* retourne toujours un singleton, nous parlons de *maxichoice contraction*. Si la fonction de *sélection* retourne l'ensemble des *remainders*, nous parlons de *full meet contraction* et notons cette fonction γ_{fc} :

$$\gamma_{fc} = \begin{array}{l} \text{Si } K \overset{R}{\perp} \varphi = \emptyset \text{ alors } \{K\} \\ \text{sinon } K \overset{R}{\perp} \varphi \end{array}$$

Un résultat important des auteurs de l'article d'AGM et d'avoir montré que la *partial meet contraction* et la *maxichoice contraction* respectent les six premiers

axiomes de contraction d'AGM alors que la *full meet contraction* respecte les huit axiomes de contraction d'AGM.

Nous allons dans les prochaines sections montrer que notre opérateur $\mathfrak{M}$ permet une construction de type *full meet contraction* et que par conséquent $\mathfrak{M}$ est un opérateur de type AGM. Nous allons d'abord présenter notre formalisation des opérateurs AGM sous l'assistant de preuve Isabelle.

V.4.2 Formalisation d'AGM sous Isabelle

Avant de démontrer l'équivalence entre l'opérateur $\mathfrak{M}$ et les opérateurs AGM, nous avons formalisé la théorie AGM sous l'assistant de preuve Isabelle. Cet outil permet à partir d'une métalogique de définir des théories afin de vérifier leur vérité et les prouver. Ces théories sont regroupées sur l'Archive of Formal Proofs (AFP) [2] et peuvent être réutilisées pour en déduire d'autres théories. Formaliser AGM sous Isabelle permet donc d'offrir tous les outils nécessaires pour permettre à quiconque de vérifier qu'un opérateur respecte l'axiomatique d'AGM. Nous allons dans cette section présenter comment nous avons découpé la formalisation d'AGM en *locales*, c'est-à-dire en modules indépendants avec leurs propres axiomes. Ce travail peut être retrouvé en détail sur l'AFP [Fouillard et al., 2021] [3].

L'ensemble des *locales* développées pour la formalisation de la théorie AGM est résumé par la figure V.2. Les flèches bleues représentent les dépendances entre les *locales* (*e.g* la *locale* de la *partial meet contraction* est une logique de Tarski). Les flèches noires montrent les équivalences entre les *locales*. Nous avons découpé la formalisation d'AGM en deux types de *locales* : les *locales descriptives*, c'est-à-dire les locales qui formalisent les axiomes d'AGM, et les *locales constructives*, c'est-à-dire les locales qui formalisent les constructions de la contraction (*i.e* les *meet contraction*). Cette séparation permet de comparer un opérateur à l'approche descriptive ou constructive d'AGM de manière indépendante.

Plus précisément, dans l'approche descriptive, nous avons développé une *locale* qui définit les six premiers axiomes de la contraction de AGM que nous nommons *AGM_Contraction*, ainsi qu'une *locale* nommée *AGM_FullContraction* qui hérite des axiomes de *AGM_Contraction* auxquels sont rajoutés les deux derniers axiomes de contraction d'AGM. Il est à noter que *AGM_Contraction* ne dépend que d'une logique de Tarski (*i.e* basé sur l'opérateur de conséquence de Tarski) alors que *AGM_FullContraction* dépend en plus d'une logique supraclassique (*i.e* basé sur les opérateurs logique classique ainsi que l'opérateur de conséquence). En plus de ces deux locales, nous définissons deux *locales* supplémentaires qui placent la contraction et la full contraction dans un contexte de logique supraclassique et compacte (respectivement *AGMC_SC* et *AGMFC_SC*). Ces deux locales sont nécessaires car l'équivalence entre les axiomes d'AGM et les *meet contractions* ne

2. https://www.isa-afp.org/
3. https://www.isa-afp.org/entries/Belief_Revision.html

peut se faire que dans ce contexte.

Dans l'approche constructive, nous suivons la même logique que l'approche descriptive. Nous avons développé trois locales pour chaque construction possible : *partial meet contraction*, *transitional relational partial meet contraciton* et *full meet contraction*. Là encore, trois autres *locales* sont définies pour placer chacune de ces constructions dans un contexte de logique supraclassique et compacte.

Nous avons pu montrer grâce à ces *locales* que dans le contexte d'une logique supraclassique et compacte, la *partial meet contraction* respecte les six premiers axiomes de la contraction d'AGM, et que la *full meet contraction* respecte l'ensemble des axiomes de la contraction d'AGM.

Maintenant que nous avons pu formaliser l'axiomatique d'AGM sous Isabelle, nous pouvons montrer que l'opérateur $\mathfrak{M}$ peut être utilisé pour construire un opérateur AGM. Pour cela nous allons d'abord montrer que calculer les MCses est équivalent à calculer les *remainders*.

V.4.3 Équivalence MCS et remainder

Afin de monter que $\mathfrak{M}$ permet de construire une *full meet contraction*, il est nécessaire de montrer que les MCS calculés par $\mathfrak{M}$ sont équivalents à des *remainders*. Nous présentons ici les différents lemmes et théorèmes que nous avons utilisés et prouvés pour arriver à cette équivalence. Nous ne nous attarderons pas sur les détails des différentes preuves, les détails pouvant être retrouvés sur un répertoire git ainsi que dans les annexes (voir Annexe A). Nous allons présenter ici le raisonnement général pour prouver l'équivalence entre les MCS et les *remainders*.

Nous montrons tout d'abord dans un premier temps le lemme $(L1)$ suivant :

$$(L1) \quad M \in \mathfrak{M}\big(K, \emptyset\big) \longleftrightarrow K \setminus M \in K \overset{R}{\perp} \perp$$

C'est-à-dire que le complément $K \setminus M$ d'un MCS M de K est un *remainder* de K éliminant faux, *i.e* un sous-ensemble maximal cohérent de K. Nous en déduisons alors le théorème $(T1)$ suivant :

$$(T1) \quad \begin{array}{l} \text{Si } M \in \mathfrak{M}\big(K, \emptyset\big) \text{ et } \varphi \in K \setminus M \\ \text{alors } K \setminus M \in K \overset{R}{\perp} \neg\varphi \end{array}$$

C'est-à-dire que si un MCS ne retire pas φ de K alors cela veut dire que le MCS correspond à un *remainder* de K pour $\neg\varphi$. En effet, $\neg\varphi$ ne peut être déduit de $K \setminus M$ du fait que $K \setminus M$ est cohérent (d'après le lemme précédent) et $\varphi \in K \setminus M$.

De ce théorème, nous formalisons les constructions qui permettent de passer d'un MCS à un *remainder* et vice-versa, ainsi :

(1) à partir d'un MCS M de $K \cup \{\neg\varphi\}$ qui ne contient pas $\neg\varphi$, on obtient $K \setminus M$ un *remainder* de K pour φ.

(2) et à partir d'un *remainder* B de K pour φ, on obtient $K \setminus B$ un MCS de $K \cup \{\neg\varphi\}$ qui ne contient pas $\neg\varphi$

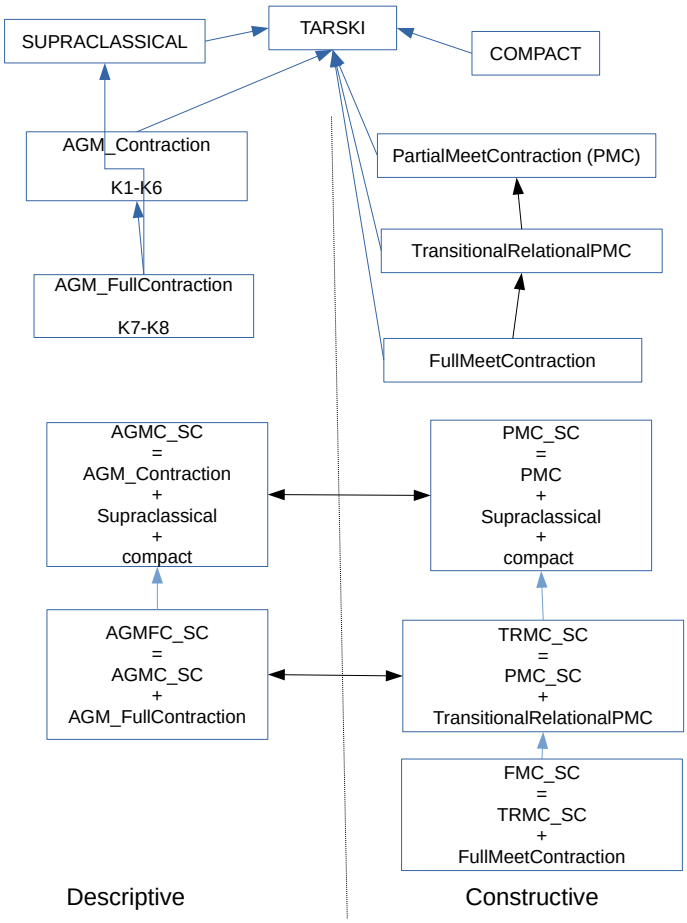


Figure V.2 – Les locales pour la formalisation d'AGM

$$(T2)\quad (M \in \mathfrak{M}\big(K \cup \{\neg\varphi\}, \emptyset\big) \wedge \neg\varphi \notin M) \Longrightarrow K \setminus M \in K \overset{R}{\perp} \varphi \qquad (1)$$
$$B \in K \overset{R}{\perp} \varphi \Longrightarrow (K \setminus B \in \mathfrak{M}\big(K \cup \{\neg\varphi\}, \emptyset\big) \wedge \neg\varphi \notin K \setminus B) \quad (2)$$

De ce théorème $(T2)$, nous concluons l'équivalence entre les MCSes et les *remainders* par le théorème $(T3)$:

$$(T3)\quad K \overset{R}{\perp} \varphi = \left\{ K - M \;\middle|\; \begin{array}{l} M \in \mathfrak{M}\big(K \cup \{\neg\varphi\}, \emptyset\big) \\ \neg\varphi \notin M \end{array} \right\}$$

Forts de cette équivalence, nous pouvons maintenant définir une fonction de *sélection* à partir des MCSes retournés par $\mathfrak{M}$ qui sélectionne l'ensemble des *remainders* et construire une *full meet contraction*.

V.4.4 Équivalence avec la full meet contraction

La *full meet contraction* pour laquelle les huit axiomes d'AGM sont respectés correspond à : $\gamma_{fc}(K \overset{R}{\perp} \varphi) = K \overset{R}{\perp} \varphi$, si $K \overset{R}{\perp} \varphi$ n'est pas vide. Or, nous savons par $(T3)$ qu'il existe une équivalence entre les *remainders* et les MCSes. Nous reprenons donc cette équivalence et notons la fonction de *sélection* $\gamma_{\mathfrak{M}}(K, \varphi)$ qui retourne l'ensemble des *remainders* de K pour φ de la manière suivante :

$$\gamma_{\mathfrak{M}}(K, \varphi) = \begin{array}{lll} \text{Si } \vdash \varphi & \text{alors} & \{K\} \\ & \text{sinon} & \left\{ K - M \;\middle|\; \begin{array}{l} M \in \mathfrak{M}\big(K \cup \{\neg\varphi\}, \emptyset\big) \\ \neg\varphi \notin M \end{array} \right\} \end{array}$$

Une fois l'équivalence $\gamma_{fc} = \gamma_{\mathfrak{M}}$ établie, nous pouvons déduire que si nous faisons l'intersection de tous les MCses retournés par l'opérateur de diagnostic $\mathfrak{M}$ alors c'est une *full contraction* $\div^{\gamma_{fc}} = \div^{\gamma_{\mathfrak{M}}}$, c'est-à-dire une contraction qui respecte les huit axiomes des contractions dans AGM. Nous pouvons aussi profiter de la *Levi identity* pour définir :

$$K *^{\gamma_{\mathfrak{M}}} \varphi = (K \div^{\gamma_{\mathfrak{M}}} \neg\varphi) + \varphi$$

comme opérateur de révision, qui vérifie par construction les axiomes des révisions AGM.

Nous signalons ici que si l'équivalence entre *remainders* et MCSes $(T3)$ a été prouvée dans le cadre d'une logique supraclassique mais pas nécessairement compacte, la conformité de l'opérateur de sélection à l'axiomatique AGM nécessite que la logique soit compacte. Très exactement, l'axiome de succès nécessite la compacité de la logique.

Dans le cas de notre algorithme, nous sélectionnons un seul MCS pour construire un état de croyances. En d'autres termes, notre algorithme de diagnostic ne sélectionne qu'un seul *remainder*. La fonction de sélection de notre algorithme correspond donc à la définition de la *maxichoice contraction*. Dans la même logique que la *full meet contraction* par la *Levi identity*, nous pouvons en conclure que notre algorithme correspond à un opérateur de révision AGM qui respecte les six premiers axiomes.

V.4.5 Conclusion

Nous avons pu montrer dans cette section que les MCS étaient équivalents à la notion de *remainders*, c'est-à-dire des sous-ensembles maximaux ne permettant pas de déduire une proposition φ. À partir de cette équivalence, nous avons pu définir une fonction de *sélection* utilisant notre opérateur de diagnostic $\mathfrak{M}$ et montrer que nous pouvions à partir de cette fonction construire une *full meet contraction* en sélectionnant l'ensemble des MCSes. Dans le cas de notre algorithme, nous sélectionnons un unique MCS retourné par $\mathfrak{M}$, cela correspond alors à une *maxichoice contraction*.

La littérature ayant déjà montré l'équivalence entre les opérateurs de révision AGM, la *full meet contraction* et la *maxichoice contraction*, nous pouvons en conclure que notre opérateur $\mathfrak{M}$ permet d'effectuer une révision de croyance à la AGM.

Nous allons dans la prochaine section présenter l'application du diagnostic avec $\mathfrak{M}$ sur le cas d'étude de l'accident du vol Rio-Paris afin d'illustrer la pertinence de cet opérateur.

V.5 Application sur le cas d'étude

Afin d'illustrer notre opérateur de diagnostic, nous allons dans cette section modéliser l'accident du 447 en nous reposant sur la description faite section IV.1. Pour cela, nous allons dans un premier temps définir les littéraux logiques que nous allons utiliser dans cette modélisation, puis dans un deuxième temps les différents éléments de la modélisation. De plus, nous détaillerons sur un pas de temps le calcul du diagnostic par notre opérateur et enfin discuterons des résultats trouvés.

V.5.1 Définition des propositions

La base de notre langage $\mathcal{L}$ repose sur un ensemble $\mathcal{P}$ de propositions indicées temporellement (voir sous-section V.1.1) qui serviront à définir les ensembles et règles nécessaires dans la modélisation. Pour les définir, nous allons nous limiter aux différents facteurs que le *Bureau d'Enquêtes et d'Analyses* a mis en avant pour leur rôle dans l'accident et dont nous avons déjà fait la liste section IV.1. Nous définissons les propositions suivantes dans $\mathcal{P}$:

- alarm_t qui indique si une alarme sonne au pas de temps t ;
- buffet_t qui indique si une vibration de type « buffeting » est ressentie ;
- $\mathrm{acceleration}_t$ qui indique si les outils de navigation indiquent une accélération au temps t ;
- stall_t qui indique si l'avion est en décrochage au temps t ;
- $\mathrm{overspeed}_t$ qui indique si l'avion est en survitesse au temps t ;
- $\mathrm{VS} \uparrow_t$ qui indique si la vitesse verticale de l'avion augmente au temps t ;

— FDpull_t qui indique que le directeur de vol indique de tirer le manche au temps t.

En plus des propositions de $\mathcal{P}$, nous définissons les actions de $\mathcal{A}$ par simplement deux propositions d'actions :

— pull_t qui indique que l'agent tire le manche de contrôle au temps t ;

— push_t qui indique que l'agent pousse le manche de contrôle au temps t.

À partir, de ces propositions, nous allons maintenant définir chaque élément de la modélisation

V.5.2 Définition des éléments du modèle

Pour définir les différents éléments du modèle, nous allons nous baser entre autre sur la description en quatre pas de temps que nous avons définie section IV.1.

Les observations Pour les observations, le pilote peut observer au premier pas de temps une accélération soudaine, une vibration et l'alarme qui sonne. De ce fait, nous avons :

$$Obs_1 \equiv \{\mathrm{buffet}_1, \mathrm{alarm}_1, \mathrm{acceleration}_1\}$$

Au deuxième pas de temps, l'alarme se désactive, les vibrations ont disparu ainsi que l'accélération. Toutefois, le pilote peut observer que le directeur de vol lui indique de tirer et que la vitesse verticale augmente. De ce fait :

$$Obs_2 \equiv \left\{ \begin{array}{l} \neg\,\mathrm{buffet}_2, \neg\,\mathrm{alarm}_2, \neg\,\mathrm{acceleration}_2, \\ \mathrm{FDpull}_2, \mathrm{VS}\uparrow_2 \end{array} \right\}$$

Le troisième pas de temps est similaire :

$$Obs_3 \equiv \left\{ \begin{array}{l} \neg\,\mathrm{buffet}_3, \neg\,\mathrm{alarm}_3, \neg\,\mathrm{acceleration}_3, \\ \mathrm{FDpull}_3, \mathrm{VS}\uparrow_3 \end{array} \right\}$$

Le quatrième et dernier pas de temps est, lui aussi, similaire à l'exception que l'alarme se réactive :

$$Obs_4 \equiv \left\{ \begin{array}{l} \neg\,\mathrm{buffet}_4, \mathrm{alarm}_4, \neg\,\mathrm{acceleration}_4, \\ \mathrm{FDpull}_4, \mathrm{VS}\uparrow_4 \end{array} \right\}$$

Les croyances initiales Nous considérons qu'au pas de temps 0, l'agent croit qu'il n'y a pas d'alarme, ni de vibrations, ni d'accélération et que l'avion n'est pas en décrochage ou en survitesse.

$$\mathcal{I}nit \equiv \left\{ \begin{array}{l} \neg\,\mathrm{alarm}_0, \neg\,\mathrm{buffet}_0, \\ \neg\,\mathrm{stall}_0, \neg\,\mathrm{overspeed}_0 \end{array} \right\}$$

Les désirs Nous considérons que le pilote ne souhaite pas être dans une situation de danger pour lui. Dans le cas de l'accident 447, les deux situations dangereuses importantes que le pilote souhaite éviter sont le décrochage et la survitesse. Le décrochage pouvant amener l'avion à s'écraser et la survitesse à ce que l'avion explose en vol. Nous avons donc :

$$\mathcal{D} \equiv \{\neg\,\mathrm{stall}_t, \neg\,\mathrm{overspeed}_t\}$$

La trace Pour les actions, nous savons que pour les deux premiers pas de temps, le pilote a décidé de tirer puis de pousser le manche au troisième pour enfin tirer à nouveau au quatrième pas de temps :

$$\mathcal{T} \equiv \{a_1 = \text{pull}_1, a_2 = \text{pull}_2, a_3 = \text{push}_3, a_4 = \text{pull}_4\}$$

Les règles de raisonnement Nous définissons les règles de raisonnement de leur ordre d'importance la plus élevée à la plus faible. Nous allons donc commencer par les règles que l'agent ne peut pas ignorer, à savoir les règles de cohérence R^C. Nous considérons qu'il existe une règle essentielle que l'agent ne peut pas ignorer :

$$R^a \equiv (\text{stall}_t \wedge \text{overspeed}_t) \rightarrow \bot$$
décrochage et survitesse s'excluent l'un l'autre

Cette règle est essentielle car elle permet au pilote de ne pas croire qu'il est dans la situation de décrochage et de survitesse en même temps, sachant que ce n'est pas possible d'un point de vue physique. Nous définissons ensuite les règles fortes R^S qui sont vraies dans la majorité des cas et sont des règles à suivre dans le cas général :

$R^b \equiv \text{buffet}_t \rightarrow \text{stall}_t$
les vibrations indiquent un décrochage

$R^c \equiv \text{alarm}_t \rightarrow \text{stall}_t$
l'alarme de décrochage indique un décrochage

$R^d \equiv (\text{VS} \uparrow_t \wedge \neg\, \text{stall}_t) \rightarrow \text{overspeed}_t$
hors décrochage, l'augmentation de la V_z correspond à une survitesse

$R^e \equiv (\text{pull}_t \wedge \text{overspeed}_t) :: \neg\, \text{overspeed}_{t+1}$
tirer le manche résout la survitesse

$R^f \equiv (\text{push}_t \wedge \text{stall}_t) :: \neg\, \text{stall}_{t+1}$
pousser le manche résout le décrochage

$R^g \equiv \{\neg\, \text{stall}_t\}\, \text{pull}_t$
tirer le manche ne doit pas se faire pendant un décrochage

$R^h \equiv \{\neg\, \text{overspeed}_t\}\, \text{push}_t$
pousser le manche ne doit pas se faire pendant une survitesse de l'avion

$R^i \equiv \text{FDpull}_t \rightarrow \text{pull}_t$
l'opérateur devrait tirer sur le manche lorsque le directeur de vol le demande

$R^j \equiv \text{acceleration}_t \rightarrow \text{overspeed}_t$
une accélération est un indicateur de survitesse

V.5.3 Exemple d'un diagnostic

Considérons que nous souhaitons déterminer les états de croyances possibles de l'agent au pas de temps 1. En partant du pas de temps 0, nous devons construire l'état de croyance du pas de temps 0, défini par l'ensemble suivant (voir section V.2) :

$$B_0 = \{\mathcal{I}nit \cup \mathcal{D} \cup \mathcal{R}\} \setminus \mathcal{I}$$

avec $\mathcal{I}$ une solution de l'opérateur de diagnostic $\mathfrak{M}$. Or l'ensemble $\{\mathcal{I}nit \cup \mathcal{D} \cup \mathcal{R}\}$ est cohérent donc $\mathfrak{M}$ ne retourne rien, car il n'y a pas besoin de correction, soit $M = \emptyset$.

Nous devons donc maintenant calculer les états de croyances possibles pour le pas de temps 1. Nous savons par l'équation V.1 que calculer un état de croyance B_1 équivaut à :

$$\begin{aligned} & B_1 = \Phi \setminus \mathcal{I} \\ \text{avec} \quad & \mathcal{I} \in \mathfrak{M}\big(\Phi, \mathit{screened}\big) \\ & \Phi = \{B_0 \cup \mathcal{D} \cup \mathit{Obs}_1 \cup a_1 \cup \mathcal{R}\} \\ & \mathit{screened} = \{a_1, R^C\} \end{aligned}$$

Ici l'opérateur $\mathfrak{M}$ ne retourne pas un ensemble, car il est nécessaire de corriger l'ensemble Φ qui est incohérent. En effet, au pas de temps 1, l'agent peut observer une alarme et une vibration qui par les règles R^b et R^c lui indiquent un décrochage. Il peut observer de plus une accélération qui lui indique une situation de survitesse (R^j), or par la règle R^a il ne peut pas croire en un décrochage et une survitesse en même temps. Il y a donc une incohérence. En appliquant l'algorithme de l'opérateur $\mathfrak{M}$, nous trouvons plusieurs solutions $\mathcal{I}_1^x$ au temps 1 :

$$\begin{aligned}
\mathcal{I}_1^a &= \{R^j(1), R^g(1), \mathcal{D}(\neg\,\mathrm{stall}_1)\} \\
\mathcal{I}_1^b &= \{\mathrm{acceleration}_1, R^g(1), \mathcal{D}(\neg\,\mathrm{stall}_1)\} \\
\mathcal{I}_1^c &= \{\mathrm{buffet}_1, \mathrm{alarm}_1, \mathcal{D}(\neg\,\mathrm{overspeed}_1)\} \\
\mathcal{I}_1^d &= \{\mathrm{buffet}_1, R^c(1), \mathcal{D}(\neg\,\mathrm{overspeed}_1)\} \\
\mathcal{I}_1^e &= \{R^b(1), \mathrm{alarm}_1, \mathcal{D}(\neg\,\mathrm{overspeed}_1)\} \\
\mathcal{I}_1^f &= \{R^b(1), R^c(1), \mathcal{D}(\neg\,\mathrm{overspeed}_1)\} \\
\mathcal{I}_1^g &= \{R^b(1), R^c(1), R^j(1)\} \\
\mathcal{I}_1^h &= \{R^b(1), R^c(1), \mathrm{acceleration}_1\} \\
\mathcal{I}_1^i &= \{R^b(1), \mathrm{alarm}_1, \mathrm{acceleration}_1\} \\
\mathcal{I}_1^j &= \{\mathrm{buffet}_1, \mathrm{alarm}_1, \mathrm{acceleration}_1\} \\
\mathcal{I}_1^k &= \{\mathrm{buffet}_1, \mathrm{alarm}_1, R^j(1)\} \\
\mathcal{I}_1^l &= \{\mathrm{buffet}_1, R^c(1), R^j(1)\} \\
\mathcal{I}_1^m &= \{\mathrm{buffet}_1, R^c(1), \mathrm{acceleration}_1\} \\
\mathcal{I}_1^n &= \{R^b(1), \mathrm{alarm}_1, R^j(1)\}
\end{aligned}$$

Il y a donc 14 états de croyances possibles au temps 1 reflétant des ignorances différentes. Chacun de ces états de croyances étant une branche de scénario pour le calcul des états de croyances suivant B_2.

Nous allons dans la prochaine section discuter des scénarios de raisonnement trouvés dans cet exemple et des rapprochements que nous pouvons faire avec l'enquête du BEA.

V.5.4 Discussion des résultats

Nous avons calculé l'ensemble des scénarios sur cette modélisation de l'accident de Rio-Paris et trouvons 903 scénarios de raisonnement. Nous avons alors effectué une analyse à la main de ces scénarios, ce qui nous a permis d'identifier trois familles de scénario :

— les scénarios dont les ignorances font ressortir l'ensemble des facteurs mis en évidence par le BEA ;
— les scénarios dont les ignorances ne correspondent pas complètement ou pas du tout au rapport du BEA mais qui pourraient malgré tout expliquer l'accident ;
— les scénarios absurdes pour lesquels les ignorances ne semblent pas traduire un comportement plausible.

V.5.4.a Scénarios conformes à l'analyse du BEA

Dans cette première famille de scénarios, nous pouvons retrouver l'ensemble des facteurs mis en évidence par le BEA et abordés section IV.1. Ces facteurs sont :

1. le buffet associé à de la survitesse ;
2. une sélectivité attentionnelle non portée sur l'alarme ;
3. une alarme considérée comme non pertinente du fait des indications du directeur de vol.

Un des scénarios que nous pouvons retrouver dans cette famille est le suivant avec $\mathcal{I}_i$ l'ensemble d'ignorances trouvé au pas de temps i pour ce scénario :

$$\begin{array}{lcl} \mathcal{I}_1 & \rightarrow & R_1^b, \mathrm{alarm}_1, \mathcal{D}(\neg\,\mathrm{overspeed}_1) \\ \mathcal{I}_2 & \rightarrow & R_1^e, \mathcal{D}(\neg\,\mathrm{overspeed}_2) \\ \mathcal{I}_3 & \rightarrow & R_2^e, R_3^h, R_3^i, \mathcal{D}(\neg\,\mathrm{overspeed}_3) \\ \mathcal{I}_4 & \rightarrow & R_4^c, \mathcal{D}(\neg\,\mathrm{overspeed}_4) \end{array}$$

Nous avons ici le pilote qui ignore au premier pas de temps l'alarme. Nous avons donc une solution qui se rapproche du deuxième facteur trouvé par le BEA du fait que l'alarme est ignorée et l'attention est portée sur une autre information : l'accélération qui elle n'est pas ignorée. Au même pas de temps, l'agent ignore aussi la règle R^b qui associe la vibration de type buffet à un décrochage. En ignorant une telle règle, nous trouvons une solution qui se rapproche du premier facteur proposé par le BEA. En effet, si le pilote n'associe plus le buffet au décrochage pour garder la cohérence avec le fait de croire à de la survitesse par l'accélération, alors il est possible qu'il associe le buffet à de la survitesse. Enfin, l'agent ignore son désir

de ne pas être en survitesse au temps 1 du fait qu'il croit être en survitesse par l'accélération de l'avion.

Pour le deuxième pas de temps, le pilote ignore la règle R^e qui lui indique que quand il est en situation de survitesse et tire sur le manche, il ne doit plus être en survitesse au pas de temps suivant. L'action n'a pas eu cet effet vu que le pilote peut observer une vitesse verticale qui augmente lui indiquant de son point de vue une situation de survitesse. De ce fait, le pilote doit considérer que son action n'a pas l'effet attendu et doit donc ignorer la règle d'effet de l'action pull pour garder la cohérence. Malgré ce non-effet de l'action, le pilote décide d'effectuer encore une fois la même action. Le BEA ne met pas en avant d'explication pour ce comportement, pourtant, nous pensons que nous pouvons faire un rapprochement avec le *biais d'engagement* [Staw, 1997] qui consiste à persister dans un même comportement avec des résultats de plus en plus négatifs.

Pour le troisième pas de temps, l'action n'a toujours pas l'effet attendu, le pilote croit qu'il est toujours en survitesse. La même règle R^e est ignorée qu'au pas de temps 2 pour les mêmes raisons. Toutefois, l'agent décide de changer de stratégie et de pousser le manche de contrôle au lieu de le tirer. Cette stratégie va à l'encontre de ce que lui demande le directeur de vol (R^i) et aussi de la précondition à l'action de pousser (R^h). Nous pensons donc que le pilote pour sortir de son *biais d'engagement* choisit donc d'ignorer ces règles afin de trouver une solution.

Pour le quatrième pas de temps, nous trouvons une solution qui se rapproche du troisième facteur proposé par le BEA. En effet, le pilote garde dans son état de croyances le fait qu'il a observé l'alarme au temps 4 mais pas la règle R^c qui fait le lien entre l'alarme et le décrochage. Cela traduit le fait que le pilote ne considère pas l'alarme comme pertinente par rapport aux autres informations comme l'indication du directeur de vol.

Nous pouvons trouver d'autres scénarios ou nous retrouvons ces facteurs, mais avec des variations dans les explications. Par exemple :

$$\begin{array}{lcl} \mathcal{I}_1 & \rightarrow & R^b_1, \mathrm{alarme}_1, \mathcal{D}(\neg\, \mathrm{overspeed}_1) \\ \mathcal{I}_2 & \rightarrow & \mathrm{VS} \uparrow_2 \\ \mathcal{I}_3 & \rightarrow & \mathrm{FDpull}_3, R^h_3, \mathcal{D}(\neg\, \mathrm{overspeed}_3) \\ \mathcal{I}_4 & \rightarrow & R^c_4, \mathcal{D}(\neg\, \mathrm{overspeed}_4) \end{array}$$

Nous retrouvons dans ce scénario les mêmes ignorances et explications correspondantes aux facteurs du BEA pour les pas de temps 1 et 4. Toutefois, au pas de temps 2 le pilote ignore simplement l'observation de l'augmentation de la vitesse verticale. Nous interprétons cette ignorance comme un *excès de confiance* : le pilote croit que son action va avoir les effets attendus et ne porte pas son attention sur des informations contradictoires. Au pas de temps 3, le pilote prend conscience qu'il est de son point de vue toujours en survitesse et décide de changer de stratégie en poussant le manche de contrôle (en ignorant R^h) et en ne portant pas son attention sur l'indication du directeur de vol (il ignore FDpull_3).

Nous pouvons donc voir que cette famille offre des scénarios qui peuvent être rapprochés des explications que nous retrouvons dans l'enquête du BEA. De plus, nous trouvons des explications supplémentaires non suggérées par le BEA qui offrent d'autres options d'explications possibles. Ces premiers résultats montrent qu'une telle formalisation de diagnostic permet de retrouver des solutions cohérentes et réalistes. Toutefois, les solutions n'ont du sens qu'à travers une interprétation du *pourquoi* des ignorances qui sont sur ces résultats faits de manière subjective et non formelle. Au-delà des solutions encourageantes trouvées par ces résultats, ceux-ci confirment l'importance d'un deuxième modèle formel dit d'*évaluation* (introduit sous-section IV.2.5) permettant d'offrir une compréhension formelle de ces ignorances par les biais cognitifs en définissant par exemple la sélectivité attentionnelle sur une proposition pour enlever toute subjectivité dans l'analyse.

V.5.4.b Scénarios différents de l'analyse du BEA

En plus des scénarios conformes aux facteurs trouvés par le BEA, nous trouvons des scénarios avec des explications différentes du BEA mais qui restent pourtant plausibles. Par exemple :

$\mathcal{I}_1$	$\rightarrow$	$\mathrm{acceleration}_1, R^g_1, \mathcal{D}(\neg\,\mathrm{stall}_1)$
$\mathcal{I}_2$	$\rightarrow$	$\mathrm{VS}\uparrow_2$
$\mathcal{I}_3$	$\rightarrow$	$\mathrm{FDpull}_3, R^h_3, \mathcal{D}(\neg\,\mathrm{overspeed}_3)$
$\mathcal{I}_4$	$\rightarrow$	$R^g_4, \mathcal{D}(\neg\,\mathrm{stall}_4)$

Nous trouvons ici qu'à l'inverse des scénarios conformes au BEA, le pilote croit qu'il est en décrochage au premier pas de temps en ne portant pas son attention sur l'accélération. Toutefois, il se trompe d'action en ignorant R^h qui indique qu'il ne doit pas tirer le manche en cas de décrochage. À première vue, l'ignorance d'une telle règle (en l'absence d'indices précédents) peut sembler forte, mais cette hypothèse n'est pas complétement écartée par le BEA du fait de « la faible exposition [...] en formation continue (théorique et pratique) au phénomène de décrochage, à l'alarme STALL » (p.196).

Au deuxième pas de temps, le pilote ne porte pas son attention sur la vitesse verticale et effectue les actions demandées par le directeur de vol (*i.e* tirer le manche).

Au troisième pas de temps, le pilote observe finalement de son point de vue qu'il est en survitesse et décide de pousser le manche alors qu'il ne doit pas du fait de la règle R^h. C'est pourquoi il ignore cette règle et ne porte pas d'attention à l'indication du directeur de vol.

Enfin, au quatrième pas de temps, le pilote croit de nouveau qu'il est en décrochage du fait de l'alarme, mais effectue la même erreur que précédemment et décide de tirer le manche de contrôle.

Nous avons donc ici un scénario dont l'explication repose essentiellement sur le manque de formation : le pilote inverse les actions à effectuer lors d'un décrochage ou d'une situation de survitesse. Ce scénario, bien que moins plausible que ceux conformes aux facteurs trouvés par le BEA, n'est pas à écarter, car il n'est pas impossible. Toutefois, la prise en compte de tels scénarios montre là encore l'importance du modèle d' *évaluation* afin de distinguer les scénarios plus ou moins plausibles en fonction de leur explication.

V.5.4.c Scénarios absurdes

Enfin, la dernière famille est composée des scénarios dits absurdes que nous considérons comme des comportements non plausibles. Par exemple :

$\mathcal{I}_1$	$\rightarrow$	$\mathrm{buffet}_1, \mathrm{acceleration}_1, \mathrm{alarm}_1$
$\mathcal{I}_2$	$\rightarrow$	$\mathrm{VS} \uparrow_2$
$\mathcal{I}_3$	$\rightarrow$	$\mathrm{FDpull}_3, \mathrm{VS} \uparrow_3$
$\mathcal{I}_4$	$\rightarrow$	$\mathrm{alarm}_4, \mathrm{VS} \uparrow_4$

Ici, le pilote ne porte aucune attention à toutes les informations importantes et de fait effectue des actions sans réelle justification. Nous ne pouvons pas alors rapprocher ces ignorances d'une explication possible. Par exemple, le pilote ne pouvant observer au temps 1, que l'accélération, l'alarme et le buffet (*i.e* $Obs_1 \equiv \{\mathrm{buffet}_1, \mathrm{alarm}_1, \mathrm{acceleration}_1\}$), il semble impossible qu'il n'ait pas observé au moins une de ces informations. De plus, l'action de tirer n'a pas de justification par les ignorances, car l'agent ne cherche pas à corriger une situation dans laquelle il se trouve (décrochage ou survitesse).

Notre opérateur de diagnostic explore un grand nombre de solutions et par conséquent remonte des solutions très peu plausibles voir impossibles. Là encore, cela montre la nécessité du modèle d'*évaluation* qui permettra de filtrer ces scénarios qui n'ont pas d'explication afin de, par exemple, ne pas explorer la branche de scénario correspondante.

V.5.5 Limites et conclusion

Nous pouvons voir que cet opérateur de diagnostic permet d'explorer un grand nombre de scénarios possibles. Ces scénarios peuvent être plus ou moins plausibles en fonction de l'interprétation des ignorances trouvées à travers des explications cognitives comme des biais. Toutefois, cette première analyse des scénarios en rapprochant les ignorances à des explications dans la littérature reste très subjective et ne repose pas sur un principe formel. De ce fait, cette analyse confirme la nécessité d'un modèle d'*évaluation* chargé de définir une plausibilité à partir de la définition formelle d'explications. Un tel modèle permettra de distinguer automatiquement les scénarios plausibles des moins plausibles et permettra en plus de filtrer les scénarios impossibles.

Toutefois, une importante limite est à considérer dans cette modélisation et ce diagnostic. Nous avons considéré ici une modélisation qui ne prend pas en compte un principe important dans la littérature de la dynamique du monde et des croyances d'un agent : *l'inertie des croyances* abordée sous-section III.1.3. Un agent doit par défaut avoir des croyances qui ne changent pas sauf changement par ses actions ou par le monde. Par conséquent, si l'agent n'a aucune information sur une proposition au temps 2, mais en avait au temps 1, alors il doit croire que par défaut cette proposition n'a pas changé depuis le temps 1. Par exemple, si nous reprenons le deuxième exemple des scénarios conformes au BEA, nous avons aux pas de temps 2 et 3 les ignorances suivantes :

$$\begin{array}{lcl} \mathcal{I}_2 & \rightarrow & \mathrm{VS}\uparrow_2 \\ \mathcal{I}_3 & \rightarrow & \mathrm{FDpull}_3, R^h_3, \mathcal{D}(\neg\,\mathrm{overspeed}_3) \end{array}$$

Au temps 2 l'agent n'ignore pas l'observation FDpull_2 mais au temps 3 l'information FDpull_3 est ignorée afin d'être cohérent avec le fait que l'agent pousse sur le manche de contrôle. Toutefois, si nous prenons en compte l'*inertie* dans la modélisation, l'agent doit croire même sans observer le directeur de vol, que celui-ci lui indique encore de tirer sur le manche car rien n'indique que par les actions de l'agent et du fait de sa non-observation du directeur de vol que celui-ci indique autre chose que précédemment. Par conséquent, l'ignorance $\mathcal{I}_3$ n'est pas suffisante en tant que telle si l'*inertie* est prise en compte, il est nécessaire d'avoir le fait que l'agent a ignoré cette *inertie* et que de ce fait il ne considère pas que le directeur de vol indique de tirer le manche comme précédemment. Il est donc nécessaire d'introduire l'*inertie* des croyances pour toutes les raisons développées chapitre III mais aussi pour considérer que des erreurs sont possibles dans cette inertie.

V.6 Conclusion

Dans ce chapitre, nous avons pu voir sous-section V.1.1 que nous basons la modélisation de l'accident sur un langage de logique propositionnelle simple avec des littéraux indicés temporellement. À partir de ce langage, nous définissons plusieurs éléments pour la modélisation avec : les observations $\mathcal{O}bs$, règles de raisonnements $\mathcal{R}$, croyances initiales $\mathcal{I}nit$, trace des actions $\mathcal{T}$ et des désirs de l'agent $\mathcal{D}$ (voir sous-section V.1.2). Tous ces ensembles permettent de définir un état de croyances à un instant t comme un ensemble B_t de propositions qui est l'union des éléments de la modélisation (voir section V.2). Nous considérons alors l'ensemble B_t comme une base de données comme l'approche de Shoham (au III.1.4.c), c'est-à-dire un ensemble ayant une contrainte de cohérence et d'unicité sur les actions. De ce fait, pour respecter la contrainte de cohérence, il est nécessaire de trouver un ensemble $\mathcal{I}$ permettant de corriger l'ensemble B_t :

$$B_t = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R}\} \setminus \mathcal{I}$$

Nous basons la sémantique de l'état de croyances d'un agent sur l'utilisation d'un SAT solveur afin de déterminer si l'agent croît, ne croît pas une proposition ou ne connait pas une proposition. Pour cela, nous utilisons des littéraux $known()$ qui permettent de marquer les propositions connues. Nous considérons que les actions, observations et croyances initiales de l'agent sont par défaut connues.

Le problème de diagnostic d'un état de croyances est alors de déterminer l'ensemble des ignorances $\mathcal{I}$ pour B_t. Nous avons montré section V.3 que cela correspondait finalement à un diagnostic très connu dans la littérature, le *consistency-based diagnosis* abordé sous-section III.2.3. Nous faisons alors appel à un opérateur $\mathfrak{M}(\Phi, screened)$ basé sur l'algorithme de [Liffiton et al., 2008] et qui recherche les MCS, c'est-à-dire les corrections minimales de Φ en immunisant à la correction le sous ensemble $screened$ de Φ. La formalisation du diagnostic est alors équivalente à l'équation V.1 où $screened$ est composé de l'action de l'agent et des règles de cohérence afin d'éviter des diagnostics inintéressants comme le fait que l'agent ne respecte pas les lois physiques ou oublie l'action qu'il est en train d'effectuer. Chaque MCS trouvé par l'opérateur $\mathfrak{M}$ permet de retrouver un état de croyances possible. Le diagnostic de l'ensemble des états de croyances pour chaque pas de temps est alors une structure en arbre du fait de la définition d'un état de croyances à partir de l'état de croyances précédent. Nous définissons alors un scénario comme un chemin dans l'arbre qui représente un diagnostic des croyances de l'agent sur l'ensemble de l'accident modélisé.

Nous avons pu montrer formellement section V.4 que l'opérateur de diagnostic $\mathfrak{M}$ respecte les axiomes d'AGM et permet donc d'effectuer une *révision minimale* au sens de AGM. Nous savons par la littérature, comme nous avons vu sous-section III.2.4, qu'un opérateur AGM est équivalent à un opérateur calculant un *consistency-based diagnosis*. De ce fait, notre opérateur à base de MCS est équivalent à un *consistency-based diagnosis*.

Enfin, nous avons proposé section V.5 une modélisation de l'accident du vol 447 abordé section IV.1. L'analyse des scénarios calculés par l'opérateur de diagnostic met en avant trois familles : des scénarios conformes à l'analyse du BEA, des scénarios différents de l'analyse du BEA mais plausibles, et des scénarios absurdes. Ces scénarios sont notamment distingués en fonction d'une explication cognitive qui correspond aux ignorances. Toutefois, ces explications sont subjectives et ne reposent pas sur une définition formelle. Cette première application vient confirmer l'importance du modèle d'*évaluation* qui recherche les explications des ignorances afin de déterminer la plausibilité d'un scénario ainsi que de retirer toute subjectivité dans l'interprétation des ignorances. En plus de la nécessité du modèle d'*évaluation*, cette première application montre aussi une limitation de la modélisation qui ne prend pas en compte l'*inertie* des croyances attendue dans un agent rationnel. En ne prenant pas en compte l'*inertie* dans la modélisation, l'opérateur de diagnostic ne capture pas toutes les ignorances de l'agent et par conséquent tous les scénarios possibles. Il est donc nécessaire pour ne pas passer à côté d'un scénario, de prendre

en compte l'*inertie* des croyances. Nous allons aborder dans le chapitre suivant quels sont les types d'ignorances possibles de l'agent en prenant en compte l'*inertie* des croyances. Nous verrons qu'il en découle un algorithme de diagnostic plus complexe, mais reposant sur les mêmes principes développés dans ce chapitre.

VI - L'inertie et la distorsion dans le modèle d'explication

Nous avons pu voir dans le chapitre précédent que bien que notre opérateur de diagnostic trouve des scénarios conformes à l'analyse du BEA sur le cas d'étude du crash du vol 447, nous n'explorons pas toutes les ignorances possibles, notamment celles liées à l'*inertie* des croyances. Par conséquent, il est possible de passer à côté de scénario plausible qui pourrait expliquer la situation. C'est pourquoi nous allons dans ce chapitre aborder comment la prise en compte de l'*inertie* introduit un nouveau problème dans le contexte d'un diagnostic. Pour cela, nous allons tout d'abord introduire un nouveau cas d'étude : l'accident du vol 5148 Air Inter sur le mont Sainte-Odile qui nous servira d'illustration pour le reste du chapitre. Nous verrons ensuite que prendre en compte l'*inertie* dans l'étude de l'erreur humaine introduit la problématique de la distorsion du décor. Nous aborderons comment nous prenons en compte l'*inertie* dans notre modèle. De plus, nous verrons qu'il existe plusieurs ignorances de natures différentes qui ne peuvent pas être toutes considérées comme des erreurs. Par la suite, nous verrons que de ces ignorances et de la prise en compte de l'*inertie* découle un algorithme de diagnostic par itération et montrerons sa correction et complétude.

VI.1 L'accident du mont Sainte-Odile

Afin d'illustrer la prise en compte de l'*inertie* dans notre modèle, nous allons nous appuyer sur un nouveau cas d'étude où une erreur liée à l'*inertie* permet d'expliquer l'accident. Ce cas d'étude est l'accident du vol 5148 de *AirInter* au mont Sainte-Odile et survenu en 1992 [BEA, 1993]. Pour cela, nous allons faire une brève description de l'accident puis présenterons une modélisation de l'accident.

VI.1.1 Déroulé de l'accident

L'enquête menée par le BEA sur l'accident du mont Sainte-Odile [BEA, 1993] montre que la raison de l'accident du vol 5148 de *AirInter* est dû à une erreur des pilotes en raison d'une interface qui peut porter à confusion. En effet, les pilotes peuvent programmer sur la même interface la vitesse de descente de l'avion (*Vertical Speed*) et l'angle descente (*Flight Path Angle*) en permutant d'un mode à l'autre par un bouton. Or dans le cas de l'accident du mont Sainte-Odile. Les pilotes ont effectué une erreur de programmation du pilote automatique avec une vitesse verticale trop grande parce qu'ils avaient oublié qu'ils avaient précédemment configuré le système en mode *Vertical Speed* au lieu de *Flight Path Angle* :

> « les hypothèses, assez probables, d'une confusion de mode vertical (résultant soit d'un oubli de changement de référence de trajectoire, soit d'une mauvaise exécution de la commande de changement) ou d'une erreur d'affichage de la valeur de consigne »(32.41).

Par conséquent, la trajectoire de l'avion correspond à une descente beaucoup trop rapide par rapport à la normale et l'appareil finira par s'écraser sur le mont Sainte-Odile.

Le BEA met en avant que cet oubli puisse être notamment expliqué par le fait que les pilotes ont porté leur attention sur la correction de la trajectoire horizontale :

> « un relâchement de l'attention de l'équipage pendant la phase de guidage radar, suivi d'une pointe instantanée de charge de travail qui l'a conduit à privilégier la navigation horizontale et l'établissement de la configuration de l'avion, et à déléguer totalement la navigation verticale aux automatismes de l'avion »(32.65)

Nous résumons alors la situation de l'accident par simplement deux pas de temps :

(t=1) Le pilote observe que sa position verticale et horizontale sont toutes les deux mauvaises. Il observe aussi que la valeur affichée est différente de 33. Enfin, il observe qu'il est en mode *Vertical Speed* et décide de la configurer à 3300 pieds par minute.

(t=2) Le pilote observe qu'il est en mode *Vertical Speed* et que l'interface affiche une valeur de 33. De plus, il observe que sa position verticale et horizontale sont toutes les deux mauvaises. Il décide de corriger sa position horizontale.

À partir de cette description de l'accident, nous allons dans les prochaines sections modéliser formellement cet accident dans notre modèle logique.

VI.1.2 Définition des propositions

Pour les littéraux de cet exemple, nous nous limitons aux différents facteurs mis en avant par le BEA que nous avons vu sous-section VI.1.1. Nous avons alors

comme propositions pour $\mathcal{P}$:

- onVS_t qui indique si l'interface est en mode *Vertical Speed* au temps t ;
- onFPA_t qui indique si l'interface est mode *Flight Path Angle* au temps t ;
- $\mathrm{display33}_t$ qui indique que l'interface affiche la valeur 33 au temps t ;
- $\mathrm{FPA33}_t$ qui indique que l'angle de descente est de 3,3 degrés ;
- $\mathrm{VS33}_t$ qui indique que la vitesse verticale est de 3 300 pieds par minute ;
- goodX_t qui indique que la position horizontale est correcte au temps t ;
- goodY_t qui indique que la position verticale est correcte au temps t ;

En plus des propositions de $\mathcal{P}$, nous définissons les actions de $\mathcal{A}$ par :

- $\mathrm{put33}_t$ qui indique que l'agent rentre la valeur 33 dans l'interface ;
- $\mathrm{controlX}_t$ qui indique que l'agent fait une manœuvre pour contrôler la position horizontale de l'appareil au temps t.

À partir de ces propositions, nous allons maintenant définir chaque élément de la modélisation

VI.1.3 Définition des éléments du modèle

Les observations Pour les observations, le pilote peut observer au premier pas de temps que l'interface est en *Vertical Speed* et que l'interface affiche une valeur différente de 33. De plus, il peut observer que la position verticale et horizontale est mauvaise. De ce fait, nous avons :

$$Obs_1 \equiv \{\neg\,\mathrm{display33}_1, \mathrm{onVS}_1, \neg\,\mathrm{goodX}_1, \neg\,\mathrm{goodY}_1\}$$

Au deuxième pas de temps, le pilote peut observer que l'interface lui affiche une valeur de 33. De plus, il peut observer les mêmes informations que précédemment :

$$Obs_2 \equiv \{\mathrm{display33}_2, \mathrm{onVS}_2, \neg\,\mathrm{goodX}_2, \neg\,\mathrm{goodY}_2\}$$

Les croyances initiales Nous considérons qu'au pas de temps 0 l'agent croit qu'il est dans une bonne position verticale et horizontale, que l'interface lui affiche une valeur différente de 33 et qu'il est en mode *Vertical Speed*. Nous avons alors :

$$\mathcal{I}nit \equiv \{\mathrm{goodX}_0, \mathrm{goodY}_0, \neg\,\mathrm{display33}_0, \mathrm{onVS}_0\}$$

Les désirs Nous considérons que le pilote souhaite être dans une bonne position pour atterrir, c'est-à-dire une bonne position verticale et horizontale ainsi qu'une vitesse verticale correcte :

$$\mathcal{D} \equiv \{\mathrm{goodX}_t, \mathrm{goodY}_t\}$$

La trace Pour les actions, nous savons que l'agent a rentré la valeur 33 dans l'interface puis décider de faire des manœuvres pour contrôler sa position horizontale. Nous avons donc :

$$\mathcal{T} \equiv \{a_1 = \mathrm{put33}_1, a_2 = \mathrm{controlX}_2\}$$

Les règles de raisonnement Pour les règles de cohérence R^C que l'agent ne peut pas ignorer, nous considérons que l'agent ne peut pas ignorer que le mode *Vertical Speed* est équivalent à ne pas être en mode *Flight Path Angle* :

$$R^a \quad \equiv \quad \mathrm{onVs}_t \longleftrightarrow \neg\, \mathrm{onFPA}_t$$

De plus, nous considérons que l'agent ne peut pas ignorer qu'en fonction du mode d'affichage dans lequel il se trouve la bonne valeur lui est affichée :

$R^b \quad \equiv \quad (\mathrm{display33}_t \wedge \mathrm{onFPA}_t) \rightarrow \mathrm{FPA33}_t$
un affichage de 33 indique un angle de 3,3 degrés en mode FPA

$R^c \quad \equiv \quad (\mathrm{display33}_t \wedge \mathrm{onVS}_t) \rightarrow \mathrm{VS33}_t$
un affichage de 33 indique une vitesse verticale
de 3300 pieds/minutes en mode VS

Nous définissons ensuite autres règles de raisonnement qui peuvent être ignorées :

$R^d \quad \equiv \quad (\mathrm{put33}_t \wedge \mathrm{onVS}_t) :: \mathrm{VS33}_{t+1}$
rentrer la valeur 33 en mode VS,
configure une vitesse de descente de 3300 pieds par minute
pour le prochain pas de temps

$R^e \quad \equiv \quad (\mathrm{put33}_t \wedge \mathrm{onFPA}_t) :: \mathrm{FPA33}_{t+1}$
rentrer la valeur 33 en mode FPA,
configure un angle de 3,3 degrés
pour le prochain pas de temps

$R^f \quad \equiv \quad \mathrm{controlX}_t :: \mathrm{goodX}_{t+1}$
contrôler sa position horizontale permet
d'avoir une bonne position horizontale au prochain pas de temps

$R^g \quad \equiv \quad \mathrm{FPA33}_t \rightarrow \mathrm{goodY}_t$
une bonne position verticale est équivalent à un angle de 3,3 degrés

$R^h \quad \equiv \quad \mathrm{VS33}_t \rightarrow \neg\, \mathrm{goodY}_t$
une vitesse verticale de 3300 pieds/minutes
entraîne une mauvaise trajectoire

VI.1.4 Conclusion

Nous avons dans cette section proposée une modélisation de l'accident du mont Sainte-Odile où un pilote a oublié le mode de configuration de l'autopilote ce qui a entrainé l'avion à descendre trop rapidement et s'écraser sur le mont Sainte-Odile. Nous allons maintenant dans les prochaines sections utiliser cet exemple

pour illustrer les différentes problématiques qu'introduit la prise en compte de l'*inertie* dans la modélisation de cet accident. Nous allons commencer par aborder la problématique dit de la distorsion du décor qui consiste à prendre en compte des erreurs dans l'*inertie* des croyances de l'agent comme une explication possible de l'accident.

VI.2 Problématique de la distorsion du décor

Nous avons vu au III.1.3.a que l'*inertie* des croyances devait être prise en compte afin que par défaut l'agent ne change pas ses croyances et ne se retrouve pas avec des croyances irrationnelles comme nous avons vu dans l'exemple YSP (au III.1.2.a) où sans inertie, l'agent peut croire qu'un pistolet se décharge tout seul. Il y a toutefois deux exceptions où les propositions peuvent "casser" cette inertie :

1. la proposition change car une action effectuée par l'agent change sa valeur de vérité ;
2. la proposition change car l'agent observe que la proposition a changé de valeur de vérité.

Le mécanisme de *mise à jour*, abordé au III.1.5.b, se charge des changements par l'action alors que le mécanisme d'*extrapolation*, abordé au III.1.5.c, se charge des changements par les observations.

Toutefois, il existe un troisième cas qui revêt un intérêt particulier pour nous : c'est lorsque l'agent fait une erreur sur l'*inertie*, et force un changement d'une proposition qui n'est pas dû à une action ou à des observations. Pour illustrer cette problématique, nous allons nous baser sur l'accident du mont Saint-Odile décrit dans la section précédente (section VI.1) dans le contexte d'un diagnostic d'une décision incohérente d'un agent.

Dans le cas de l'accident du mont Sainte-Odile, au premier pas de temps, nous nous retrouvons avec une décision incohérente avec le désir de l'agent d'avoir une vitesse verticale correcte. En effet, le pilote peut observer qu'il est en mode *Vertical Speed*, mais décide de rentrer la valeur 33 dans l'autopilote. Cette action a pour conséquence que la vitesse verticale soit de 3300 pieds/minutes au pas de temps suivant, ce qui est beaucoup trop haut. De ce fait, le désir du pilote d'avoir une vitesse correcte n'est pas satisfait. Si nous recherchons une explication possible de cette décision incohérente, nous pouvons trouver comme solution le pilote n'a pas observé qu'il était en mode *Vertical Speed*. Cependant, cette solution n'est pas suffisante pour expliquer la décision du pilote si nous prenons en compte l'*inertie* des croyances du pilote.

En effet, si nous considérons que la modélisation de l'accident prend en compte l'*inertie*, alors cela veut dire par exemple que la proposition onVS_t ne change pas

par défaut, sauf s'il y a un changement par une action ou une observation. Par conséquent, du fait que le pilote croit qu'il est en mode *Vertical Speed* au temps initial (*i.e* $\mathrm{onVS}_0 \in \mathcal{I}nit$), que l'agent n'effectue pas une action pour changer le mode de l'autopilote et que l'agent peut observer que le mode n'a pas changé, celui-ci doit croire par défaut qu'au pas de temps 1 il est toujours en mode *Vertical Speed*. En d'autres termes, $\mathrm{onVS}_0 \longleftrightarrow \mathrm{onVS}_1$. Par conséquent, la seule ignorance de l'observation de onVS_1 pour expliquer la décision de l'agent n'est pas suffisante, car par *inertie*, l'agent peut déduire onVS_1. Pour que la décision du pilote soit cohérente, il est donc nécessaire aussi d'ignorer l'*inertie* sur le mode *Vertical Speed* de l'autopilote de l'avion pour considérer que le pilote a oublié dans quel mode il était.

Ainsi, dans un contexte de diagnostic de décision incohérente, il est nécessaire de prendre en compte des changements non attendus dans l'*inertie* des croyances qui permettent d'expliquer un comportement incohérent. C'est-à-dire des changements dans l'*inertie*, que nous nommons *distorsions* qui ne sont pas dues aux mécanismes de *mise à jour* ou d'*extrapolation*. Nous définissons alors la prise en compte de ces *distorsions* comme la problématique de la distorsion du décor.

Nous allons dans les prochaines sections aborder comme nous introduisons l'*inertie* des croyances dans notre modèle. Nous verrons ensuite comment nous déterminons les *distorsions* qui permettent d'expliquer des décision incohérentes.

VI.3 Prise en compte de l'inertie des croyances

Avant de considérer de capturer les *distorsions* dans l'*inertie* des croyances, nous devons prendre en compte l'*inertie* dans notre modèle, c'est-à-dire que par défaut, si l'agent croit que φ_{t-1} au pas de temps t alors il doit croire que φ_t au pas de temps t. Dans notre cas, nous avons vu sous-section V.2.1 que croire en φ_t à un instant t (*i.e* $B_t \vdash \varphi_t$) correspond à deux choses :

1. $\neg\varphi_t$ est insatisfaisable dans l'état B_t
2. $\neg known(\varphi)_t$ est insatisfaisable dans l'état B_t

En d'autres termes, un littéral φ_t est crûe si elle ne peut être que vraie et connue dans l'état B_t. Ainsi l'*inertie* doit se faire à la fois sur la valeur de vérité d'un littéral mais aussi sur la connaissance de ce littéral. En effet, si l'agent sait que φ_{t-1} est vrai au pas de temps t, par défaut il doit savoir que φ_t au pas de temps t.

Nous allons dans cette section présentée la solution que nous avons choisie pour prendre en compte l'*inertie* des croyances. Enfin, nous verrons à partir de quel moment et sur quels littéraux nous introduisons l'*inertie* dans notre modèle.

VI.3.1 L'inertie : comment ?

Afin de prendre en compte l'*inertie* à la fois sur la valeur de vérité des littéraux et la connaissance des littéraux, nous allons introduire de nouvelles propositions qui forcent des littéraux à garder la même valeur entre deux pas de temps. Nous notons :

— $keep_t^{(val)}(\varphi) \equiv \varphi_{t-1} \longleftrightarrow \varphi_t$ qui indique que le littéral φ_t garde la même valeur de vérité entre le pas de temps $t-1$ et t ;

— $keep_t^{(known)}(\varphi) \equiv known(\varphi_{t-1}) \longleftrightarrow known(\varphi_t)$ qui indique que le littéral φ_t garde le même état de connaissance entre le pas de temps $t-1$ et t.

où φ est un symbole de proposition (*i.e* $\varphi \in \mathcal{S}$) correspondant au symbole du littéral dont l'*inertie* doit être prise en compte. L'indice t correspond au pas de temps pour lequel le littéral doit avoir une *inertie* par rapport au pas de temps précédent.

Les propositions $keep_t^{(known)}(\varphi)$ sont nécessaires, car s'il n'y a pas d'*inertie* sur les propositions *known* alors l'agent peut ne pas déduire d'autres croyances à partir des croyances avec une *inertie*. Par exemple, si nous avons le système logique :

$$\Phi \quad = \quad \mathrm{VS33}_1 \wedge known(\mathrm{VS33})_1 \wedge (R_1^h \equiv \mathrm{VS33}_1 \rightarrow \neg\, \mathrm{goodVS}_1)$$

Nous savons que pour des raisons de sémantique (voir sous-section V.2.1), la règle R_1^h équivaut à :

$$R_1^h \equiv (\mathrm{VS33}_1 \wedge known(\mathrm{VS33})_1) \rightarrow (\neg\, \mathrm{goodVS}_1 \wedge known(\mathrm{goodVS})_1)$$

L'agent doit connaitre $\mathrm{VS33}_1$ pour que $\neg\, \mathrm{goodVS}_1$ soit déduit et connu par la règle R_1^h. Ainsi si nous ajoutons uniquement $keep_2^{(val)}(\mathrm{VS33})$ dans le système Φ alors nous avons forcément onVS_2 à vrai mais la valeur de $known(\mathrm{VS33})_2$ est laissée libre. Par conséquent, $\neg\, \mathrm{goodVS}_2$ n'est pas forcé à vrai par la règle R_2^h. C'est pourquoi pour prendre en compte l'*inertie* des croyances, il est nécessaire aussi que l'*inertie* se fasse sur les propositions *known* afin que l'agent puisse déduire de nouvelles croyances à partir des croyances avec une *inertie*.

Ces littéraux permettent donc de garder une *inertie* entre un pas de temps $t-1$ et t de n'importe quel littéral du langage $\mathcal{L}$ défini sous-section V.1.1. Utiliser des propositions pour spécifier directement l'*inertie* des croyances a l'avantage de pouvoir considérer des ignorances possibles sur un littéral à un pas de temps précis. Par exemple, si $keep_2^{(val)}(onVS)$ est ignoré, cela veut dire que l'agent a changé sa croyance sur le mode de l'autopilote au pas de temps 2 par rapport au pas de temps 1. De ce fait, nous pouvons repérer tout changement que ce soit dû à une *mise à jour*, *extrapolation* ou *distorsion* (voir section VI.2).

Toutefois, se pose la question de quand et sur quels littéraux sont introduits ces propositions d'*inertie* ? À partir de quel moment les propositions $keep$ sont nécessaires ? Nous allons aborder ces problématiques dans la section suivante.

VI.3.2 L'inertie : où et quand ?

Nous avons vu au III.1.3.a que pour prendre en compte le *problème du décor* et par conséquent l'*inertie*, toutes les croyances d'un agent doivent avoir une *inertie* par défaut. À partir de là, une première solution naïve serait d'introduire pour toutes croyances φ de l'agent à un instant t (*i.e* $\forall\varphi$ tq $B_t \vdash \varphi$) les propositions $keep_t^{(val)}(\varphi)$ et $keep_t^{(known)}(\varphi)$ correspondantes. Or cela n'est pas nécessaire.

En effet, nous avons vu sous-section V.2.2 que toute croyance φ de B_t est soit :

— une croyance initiale ou une observation ;

— une croyance qui découle d'une action effectuée, des observations ou des croyances initiales par les règles de raisonnement $\mathcal{R}$.

Par conséquent, les croyances φ déduites par $\mathcal{R}$ pour le pas de temps $t-1$ seront aussi déduites pour le pas de temps t si l'*inertie* est introduite pour les croyances initiales, observations et effets directs d'action. Par exemple, si nous avons :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\mathrm{VS33}_1\} \\ Obs_2 & \equiv & \emptyset \\ \mathcal{R} & \equiv & \left\{ \; R^h \equiv \mathrm{VS33}_t \rightarrow \neg\,\mathrm{goodVS}_t \; \right\} \end{array}$$

Nous avons $B_1 \vdash \mathrm{VS33}_1$ et $B_1 \vdash \neg\,\mathrm{goodVS}_1$. En ajoutant seulement dans B_2 les propositions $keep_2^{(val)}(\mathrm{VS33})$ et $keep_2^{(known)}(\mathrm{VS33})$ nous avons bien $B_2 \vdash \mathrm{VS33}_2$ et $B_2 \vdash \neg\,\mathrm{goodVS}_2$ avec seulement une *inertie* sur l'observation de la VS. De ce fait, les propositions $keep$ pour un instant t correspondent à ajouter les propositions $keep$ des observations de l'instant t, des effets directs de l'action effectuée en plus des propositions $keep$ de l'instant $t-1$ mise à jour pour l'instant t. Autrement dit, il suffit d'ajouter au temps t les $keep$ des φ pour lesquels les $keep$ de φ du temps $t-1$étaient vrais. Les autres seront déduits.

Formellement, nous construisons récursivement l'ensemble $\overrightarrow{K}_t$ des littéraux initiaux et/ou observés et/ou conséquences d'une action effectuée depuis le début de la simulation :

$$\overrightarrow{K}_t = \bigcup_{\varphi_{t'} \in Obs_t \cup effect(a_t)} \{\varphi_{t'}\} \cup \overrightarrow{K}_{t-1}$$

$$\overrightarrow{K}_0 = \bigcup_{\varphi_{t'} \in \mathcal{I}nit} \{\varphi_{t'}\}$$

Il suffit d'ajouter des $keep$ pour les éléments de K pour garantir l'inertie (les autres littéraux étant déduits à partir des règles du modèle). Cependant, nous devons prendre en compte trois cas selon la valeur de l'indice temporel t' du littéral $\varphi'_t \in \overrightarrow{K}_t$:

— Si $t' = t$ alors cela veut dire que l'agent n'a pas besoin à l'instant t d'*inertie* sur φ car $B_t \vdash \varphi_t$.

— Si $t' > t$ alors cela veut dire que l'agent déduit une croyance dans le futur. Par exemple, il peut observer au temps 1 qu'il y aura des nuages au temps 3. L'*inertie* sur cette information ne doit être introduite alors qu'à partir du pas de temps 3 du fait que l'agent n'a aucune information sur l'état des nuages avant le pas de temps 3. Par conséquent, si $t' > t$ les propositions $keep$ ne sont pas ajoutées.
— Si $t' < t$ alors cela veut dire que l'agent a l'information de l'état de φ dans le passé. Par conséquent, nous devons introduire l'*inertie* du pas de temps t' au temps t. En effet, par exemple, si l'agent avait observé des nuages au temps 1 alors au temps 3 l'inertie de cette information doit être prise en compte, c'est-à-dire que du pas de temps 1 au pas de temps 3 la croyance sur les nuages doit être la même que l'instant précédent.

Formellement, nous notons l'ensemble $\mathfrak{K}_t^{(val)}$ et $\mathfrak{K}_t^{(known)}$ les ensembles qui respectivement contiennent toutes les propositions $keep_t^{(val)}(\varphi)$ et $keep_t^{(known)}(\varphi)$ qui doivent être ajoutées à l'instant t. Nous les définissons par :

$$\begin{aligned}
\mathfrak{K}_t^{(val)} &= \bigcup_{\varphi_{t'} \in \vec{K}_t} \bigcup_{i=t'}^{t} \left\{ keep_i^{(val)}(\varphi) \right\} \text{ tq } t' < t \\
\mathfrak{K}_t^{(known)} &= \bigcup_{\varphi_{t'} \in \vec{K}_t} \bigcup_{i=t'}^{t} \left\{ keep_i^{(known)}(\varphi) \right\} \text{ tq } t' < t
\end{aligned}$$

Nous notons l'union de ces deux ensembles :

$$\mathfrak{K}_t = \mathfrak{K}_t^{(val)} \cup \mathfrak{K}_t^{(known)}$$

Ainsi pour prendre en compte l'*inertie* à un instant t nous devons ajouter l'ensemble $\mathfrak{K}_t$. La nouvelle définition d'un état de croyances devient alors :

$$B_t = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R} \cup \mathfrak{K}_t\} \setminus \mathcal{I}$$

VI.3.3 Conclusion

Nous prenons en compte l'*inertie* des croyances en ajoutant un ensemble $\mathfrak{K}_t$ de littéraux $keep_t^{(val)}(\varphi)$ et $keep_t^{(known)}(\varphi)$ qui force la valeur de vérité et la connaissance d'une croyance φ à rester similaire par rapport à l'instant précédent. Ces littéraux ne sont ajoutés que pour les croyances initiales, les observations et les effets directs d'actions car l'*inertie* découle naturellement de ceux-ci.

En ajoutant $\mathfrak{K}_t$ à la définition de B_t nous permettons que des propositions d'*inertie* soient ignorées par $\mathcal{I}$ pour retrouver la cohérence. Toutefois, nous allons voir dans la prochaine section que ces ignorances ne sont pas tous de mêmes natures et ne représentent pas la même chose.

VI.4 De nouvelles incohérences liées à l'inertie, des ignorances toujours différentes mais toujours un même opérateur

Nous avons pu voir dans le chapitre précédent section V.3 que le processus de diagnostic pour un état de croyances devait faire face à deux types d'incohérences :

— des incohérences avec l'action effectuée ;

— des incohérences avec les observations.

Ces deux incohérences correspondent à respectivement un problème de *consistency-based diagnosis* et de *révision de croyance* dont les solutions peuvent être tous deux calculés par un même opérateur $\mathfrak{M}$ que nous avons défini sous-section V.3.2. Toutefois, en introduisant l'*inertie* dans la définition de l'état de croyances, nous introduisons aussi un troisième type d'incohérence : les incohérences liées à l'*inertie*.

Nous allons dans cette section aborder les différentes sous problématiques de l'incohérence dans l'*inertie* et les relier aux problèmes que nous trouvons dans la littérature. Nous verrons qu'il en ressort des ignorances de natures différentes.

VI.4.1 Deux sous problèmes d'incohérences

Tout comme section V.3, nous distinguons deux types d'incohérence liés à l'*inertie* :

1. Des incohérences d'*inertie* avec les observations, c'est-à-dire que l'agent observe des informations en contradiction avec l'*inertie* des croyances précédentes.
2. Des incohérences d'*inertie* avec l'action, c'est-à-dire que l'action rentre en contradiction avec l'*inertie* des croyances précédentes.

Nous allons illustrer chaque type d'incohérence par des exemples dans les prochaines section en les rattachant à la littérature. Enfin, nous argumenterons sur l'utilisation de notre opérateur de diagnostic $\mathfrak{M}$ pour trouver des solutions à ces incohérences.

VI.4.1.a Incohérence d'inertie et observations

Dans le cas d'une incohérence d'*inertie* liée aux observations, nous tombons dans le cas de la prise en compte des changements implicites des croyances dues aux observations. Par exemple, considérons un agent qui croit qu'il a une bonne vitesse verticale au pas de temps initial puis observe une vitesse verticale trop élevée au pas de temps suivant :

$$\begin{array}{lcl} \mathcal{I}nit & \equiv & \{\mathrm{goodVS}_0\} \\ Obs_1 & \equiv & \{\mathrm{VS33}_1\} \\ \mathcal{R} & \equiv & \{\ R^h \ \equiv \ \mathrm{VS33}_t \rightarrow \neg\,\mathrm{goodVS}_t\ \} \end{array}$$

En prenant en compte l'inertie, nous devons rajouter :

$$\mathfrak{K}_1 \ \equiv \ \{keep_1^{(val)}(\mathrm{goodVS}), keep_1^{(known)}(\mathrm{goodVS})\}$$

Nous avons alors une incohérence, car $\mathrm{goodVS}_0 \longleftrightarrow \mathrm{goodVS}_1$ et $\mathrm{VS33}_1 \rightarrow \neg\,\mathrm{goodVS}_1$. Or cela correspond à une situation où l'agent a une croyance ancienne sur le monde et doit la mettre à jour, car le monde a évolué entre temps. Ce qui correspond à la définition de l'*extrapolation* abordé au III.1.5.c.

Ce mécanisme peut être calculé par un opérateur de *révision* de type AGM sur les littéraux d'*inertie*. C'est-à-dire que les ignorances de l'*extrapolation* correspondent uniquement au $\mathfrak{K}_t^{(val)}$. Dans l'exemple ci-dessus, la solution à l'*extrapolation* est d'ignorer $keep_t^{(val)}(\varphi)$. Notre opérateur de diagnostic $\mathfrak{M}$ étant un opérateur AGM (voir section V.4), nous pouvons l'utiliser pour trouver les ignorances de l'*extrapolation*.

Il est à noter que l'*extrapolation* est un processus attendu dans un agent rationnel. Par conséquent, les ignorances trouvées par ce mécanisme ne correspondent pas à des erreurs de la part de l'agent. Il sera donc nécessaire de différencier ces ignorances dans le modèle d' *évaluation* et de les considérer comme "rationnel".

VI.4.1.b Incohérence d'inertie avec l'action

Dans le cas d'une incohérence d'inertie avec l'action, nous devons distinguer deux cas avec des ignorances de nature différente :

- Des ignorances qui sont attendues chez un agent rationnel. Par exemple, suite à l'action de l'agent de changer la valeur de l'autopilote, celui-ci doit croire que l'inertie sur la valeur de l'autopilote a changé.
- Des ignorances qui ne sont pas attendues chez un agent rationnel. Par exemple, l'agent croit que la valeur de l'autopilote a changé sans qu'il puisse observer un changement ou effectuer une action qui change la valeur de l'autopilote.

Nous allons aborder plus précisément cette distinction et la rattacher aux différents problèmes que nous avons aborder dans la littérature et les travaux de cette thèse.

Les ignorances attendues Considérons l'exemple d'un agent qui décide de rentrer la valeur 33 dans l'autopilote lorsqu'il est en mode *Vertical Speed*, ce qui a pour effet d'avoir une vitesse verticale de 3300 pieds par minutes le pas de temps suivant :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\mathrm{onVS}_1, \neg\,\mathrm{VS33}_1\} \\ \mathcal{R} & \equiv & \{\ R^e \ \equiv \ (\mathrm{put33}_t \wedge \mathrm{onVS}_t) :: \mathrm{VS33}_{t+1}\ \} \\ \mathcal{T} & \equiv & \{\mathrm{put33}_1\} \end{array}$$

avec pour inertie au temps 2 :

$$\mathfrak{K}_2 \equiv \left\{ \begin{array}{l} keep_2^{(val)}(\mathrm{onVS}), keep_2^{(known)}(\mathrm{onVS}) \\ keep_2^{(val)}(\mathrm{VS33}), keep_2^{(known)}(\mathrm{VS33}) \end{array} \right\}$$

Le système est incohérent, car nous avons $\mathrm{VS33}_2$ par la règle R^e et $\neg\,\mathrm{VS33}_2$ par $\mathfrak{K}_2$. Or, nous sommes dans une situation où l'agent modifie le monde par son action, il doit donc mettre à jour ses croyances en conséquence. Ce mécanisme correspond à la *mise à jour* que nous avons abordé au III.1.5.b. Ce mécanisme peut être donc calculé grâce à un opérateur KM de *mise à jour minimale*. Par exemple, la solution attendue dans l'exemple ci-dessus est que $keep_2^{(val)}(\mathrm{VS33})$ soit ignoré afin que l'agent considère que la vitesse verticale a changé après avoir configuré l'autopilote.

Tout comme l'*extrapolation*, la *mise à jour* retourne des ignorances attendues dans un agent rationnel. L'agent prend en compte les changements qu'il effectue sur le monde. Le modèle d'*évaluation* devra donc pouvoir distinguer ces ignorances et les considérer comme rationnelles.

Les ignorances non attendues Un deuxième cas d'incohérence d'*inertie* avec l'action est le fait d'avoir une action incorrecte quand l'*inertie* est prise en compte. C'est-le-cas dans l'exemple utilisé pour le problème de la distorsion du décor section VI.2 : l'agent ne souhaitait pas que sa vitesse soit trop élevée, mais a pourtant rentré une valeur trop élevée dans l'autopilote alors qu'il aurait dû savoir qu'il était en mode *Vertical Speed* par l'*inertie*. Le problème de la distorsion du décor est donc un problème de diagnostic de cohérence qui prend en compte l'*inertie* : nous recherchons les ignorances qui permettent de retrouver la cohérence dans les croyances de l'agent et qui permettent d'expliquer son action en considérant que celui-ci peut ignorer l'*inertie* de certaines croyances. Par conséquent, notre opérateur de diagnostic $\mathfrak{M}$ qui effectue un *consistency-based diagnosis* peut être utilisé pour trouver ces ignorances.

VI.4.2 Un même opérateur

Nous avons vu que prendre en compte l'*inertie* ajoute de nouvelles incohérences à prendre en compte qui sont liées à l'*extrapolation*, la *mise à jour* et la *distorsion du décor*. Nous avons aussi vu que l'*extrapolation* et la *distorsion du décor* peuvent être résolu par notre opérateur de diagnostic $\mathfrak{M}$ bien que les ignorances soient de natures différentes. Toutefois, la *mise à jour* ne correspond pas à un problème de *révision* ou de *consistency-based diagnosis* et à première vue ne peut pas être capturé par notre opérateur $\mathfrak{M}$. Nous allons voir que la réponse n'est pas si simple.

En effet, nous avons pu voir sous-section III.1.3 qu'une première solution proposée pour prendre en compte la *mise à jour* des propositions suite à une action et d'effectuer une *circumscription* [McCarthy, 1986]. C'est-à-dire à sélectionner les modèles qui minimisent le nombre de changements. Or la *circumscription* peut

être réduit à une *révision de croyance* à la AGM [Liberatore et al., 1997]. De ce fait, nous pouvons utiliser notre opérateur $\mathfrak{M}$ pour calculer une *circumscription*. Néanmoins, la solution d'utiliser la *circumscription* n'a pas été retenue dans la littérature du fait que les changements dans les modèles retournés pouvaient être des changements qui ne sont pas attendus par rapport à une *mise à jour* rationnel (*e.g* le problème YSP au III.1.3.a). Toutefois, dans notre cas, nous avons vu que ces changements non attendus correspondent à des *distorsions* que nous souhaitons capturer (voir section VI.2). Les solutions de la *circumscription* qui étaient donc un problème pour la *mise à jour* sont attendues dans notre cadre d'application.

Nous pouvons donc utiliser notre opérateur $\mathfrak{M}$ pour capturer à la fois les ignorances de *mise à jour*, de *distorsion*, d'*extrapolation* et de *révision de croyance*.

VI.4.3 Conclusion

Prendre en compte l'*inertie* introduit de nombreuses incohérences relatives à la *mise à jour*, *l'extrapolation* et la *distorsion*. Les ignorances qui résultent de ces mécanismes sont de natures différentes. La *mise à jour* et l'*extrapolation* retournent des ignorances rationnelles, c'est-à-dire des ignorances sur les littéraux d'*inertie* afin que l'agent prenne en compte les changements dûs à une action ou à l'évolution du monde. Les ignorances dûs à la *distorsion* sont des erreurs dans l'*inertie* des croyances de l'agent.

Nous avons pu voir que notre opérateur de diagnostic $\mathfrak{M}$ peut être utilisé pour résoudre ces trois types d'incohérences car trouver les solutions de *l'extrapolation*, *la mise à jour* et la *distorsion* se réduit à trouver les solutions d'une *révision de croyance minimale* à la AGM.

Nous allons dans la prochaine section définir notre nouvel algorithme de diagnostic permettant de distinguer et de déterminer ces différentes ignorances de natures différentes.

VI.5 Un algorithme par itération

Nous avons pu voir sous-section V.3.1 et section VI.4 que dans notre problématique de diagnostic d'un état de croyances à un instant t doit faire face à des ignorances de nature différente :

1. une préférence sur les croyances face à des informations contradictoires (*révision*) ;
2. une erreur de raisonnement (*consistency-based diagnosis*) ;
3. une mise à jour des croyances face à l'évolution du monde (*extrapolation*) ;
4. une mise à jour des croyances face à une action effectuée (*mise à jour*) ;
5. une erreur dans l'inertie des croyances (*distorsion*).

Ces ignorances étant de natures différentes, d'un point de vue purement diagnostic il est important de pouvoir les distinguer afin de comprendre au mieux l'état de croyances de l'agent. Nous avons pu voir sous-section V.3.1 et section VI.4 que notre opérateur de diagnostic $\mathfrak{M}$ permet de trouver les ignorances de chacune de ces problématiques. Nous allons donc nous baser sur cet opérateur afin de créer un nouvel algorithme de diagnostic permettant de distinguer ces différentes ignorances.

Toutefois, si nous calculons un état de croyances en prenant en compte l'*inertie* simplement en appliquant le même procédé que l'équation V.1, nous avons :

$$\begin{aligned} & B_t = \Phi \setminus \mathcal{I} \\ \text{avec} \quad & \mathcal{I} \in \mathfrak{M}\big(\Phi, screened\big) \\ & \Phi = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R} \cup \mathfrak{K}_t\} \\ & screened = \{a_t, R^C\} \end{aligned}$$

Ce qui nous retourne l'ensemble des ignorances de nature (1-5) dans un même ensemble $\mathcal{I}$ sans qu'il y ait une distinction de la nature des ignorances. Nous proposons dans cette section un algorithme par itération qui permet de calculer les ignorances d'un pas de temps t pour un problème à la fois. C'est-à-dire de calculer un premier MCS pour le problème de la *révision*, puis un MCS pour le problème du *consistency-based diagnosis* et ainsi de suite. Ainsi ce n'est pas un unique MCS M qui permet de corriger B_t mais une suite de MCSes où chaque MCS représente des ignorances de natures distinctes et ainsi donner des indices sur le pourquoi de l'ignorance pour le modèle d' *évaluation*.

Nous allons dans cette section voir un par un les différents MCSes calculés et voir comment nous utilisons cette suite de MCSes pour déterminer $\mathcal{I}$ et B_t.

VI.5.1 Ignorances liées à la révision de croyance

La première étape de notre algorithme est de déterminer les ignorances qui résultent d'une préférence sur les croyances face à des informations contradictoires, c'est-à-dire un problème de *révision de croyance* (voir au III.1.5.a). En d'autres termes, l'incohérence dans le système provient des observations que l'agent peut prendre en compte et non de son action ou de l'*inertie* des croyances. Le principe est donc dans cette première étape de ne pas considérer l'action et les littéraux d'*inertie* pour déterminer les incohérences avec les observations. Nous notons M_{rev} un MCS possible de cette première étape et B_t^{rev} l'état de croyances résultant de M_{rev}. Le calcul des ignorances relatif à la *révision de croyance* correspond alors à :

$$\begin{aligned} & B_t^{rev} = \Phi \setminus M_{rev} \\ \text{avec} \quad & M_{rev} \in \mathfrak{M}\big(\Phi, screened\big) \\ & \Phi = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \mathcal{R}\} \\ & screened = \{R^C\} \end{aligned} \tag{VI.1}$$

Pour illustrer les différentes étapes de notre algorithme, nous allons utiliser la modélisation de l'accident du mont Sainte-Odile vu section VI.1. Dans le cas

de cet accident, le système $\{\mathcal{I}nit \cup \mathcal{D} \cup \mathcal{R}\}$ est cohérent, par conséquent $B_0 = \{\mathcal{I}nit \cup \mathcal{D} \cup \mathcal{R}\}$ car il n'y a pas besoin d'ignorance pour corriger le système. Considérons que nous souhaitons maintenant calculer B_1 par notre algorithme par itération. La première étape est donc de calculer B_1^{rev} pour déterminer les incohérences liées à la *révision de croyance*. Or le système $\Phi = \{B_0 \cup \mathcal{D} \cup Obs_t \cup \mathcal{R}\}$ est incohérent du fait que l'agent observe qu'il n'est pas dans la bonne position (verticale et horizontale) alors qu'il le désire. Nous avons alors comme possibilité pour M_{rev}, les MCSes suivants que nous notons M_{rev}^x :

$$\begin{array}{lcl} M_{rev}^a & = & \{\mathcal{D}(\mathrm{goodXaxis}_1), \mathcal{D}(\mathrm{goodYaxis}_1)\} \\ M_{rev}^b & = & \{\mathcal{D}(\mathrm{goodXaxis}_1), \neg\, \mathrm{goodYaxis}_1)\} \\ M_{rev}^c & = & \{\neg\, \mathrm{goodXaxis}_1, \neg\, \mathrm{goodYaxis}_1\} \\ M_{rev}^d & = & \{\neg\, \mathrm{goodXaxis}_1, \mathcal{D}(\mathrm{goodYaxis}_1)\} \end{array}$$

En d'autres termes, l'agent peut effectuer une *révision de croyance* pour prendre en compte qu'il n'est pas en bonne position et rejeter ses désirs de l'être, ou au contraire rejeter les informations sur sa position et garder les désirs.

Chaque M_{rev}^x correspond à un état de croyance B_1^{rev} possible. Ainsi la deuxième étape qui s'occupe des ignorances qui résultent d'une action incohérente doit être calculé pour chaque état B_1^{rev} possible. Nous abordons cette deuxième étape dans la section suivante.

VI.5.2 Ignorances liées à un diagnostic

La deuxième étape de notre algorithme calcule les ignorances liées à une erreur de raisonnement, c'est-à-dire une ignorance qui résulte d'une action incohérente par rapport aux croyances et observations de l'agent à l'instant t sans considérer l'*inertie* des croyances. Cela correspond à rechercher une explication à une action effectuée et par conséquent à un *consistency-based diagnosis* comme nous avons vu section V.3. Nous allons donc reprendre exactement la même solution que l'équation V.1 en ne considérant pas l'*inertie* dans l'état de croyances de l'agent. Nous notons M_{diag} un MCS possible de cette deuxième étape et B_t^{diag} l'état de croyances résultant de M_{diag}. Le calcul des ignorances relatif au *consistency-based diagnosis* correspond alors à :

$$\begin{array}{ll} & B_t^{diag} = \Phi \setminus M_{diag} \\ \text{avec} & M_{diag} \in \mathfrak{M}\big(\Phi, screened\big) \\ & \Phi = \{B_t^{rev} \cup a_t\} \\ & screened = \{a_t, R^C\} \end{array} \qquad \text{(VI.2)}$$

où B_t^{rev} est un état de croyance résultant de la première étape de l'algorithme (voir équation VI.1).

Reprenons l'exemple de l'accident du mont Saint-Odile et considérons que nous souhaitons calculer la deuxième étape de notre algorithme à partir de l'état B_1^{rev} qui

résulte du MCS $M^a_{rev} = \{\mathcal{D}(\mathrm{goodXaxis}_1), \mathcal{D}(\mathrm{goodYaxis}_1)\}$ de l'étape précédente (voir l'exemple sous-section VI.5.1). C'est-à-dire que l'agent a décidé de prendre en compte qu'il n'était pas dans la bonne position au pas de temps 1. À partir de là, si nous recherchons B_t^{diag} avec le système $\Phi = \{B_t^{rev} \cup a_t\}$ alors, nous avons une incohérence. En effet, l'agent peut observer qu'il est en mode VS et fait l'action $\mathrm{put33}_1$. Par conséquent, par la règle R^d et la règle R^h, l'agent peut en conclure que par son action, il n'aura pas une bonne vitesse verticale au temps 2, ce qui est incohérent avec son désir. Nous avons alors différentes possibilités pour M_{diag} :

$$\begin{array}{rcl} M^a_{diag} & = & \{\mathrm{onVS}_1\} \\ M^b_{diag} & = & \{\mathcal{D}(\mathrm{goodVS}_2)\} \\ M^c_{diag} & = & \{R^d(1)\} \\ M^d_{diag} & = & \{R^h(2)\} \end{array}$$

Ces différentes ignorances représentent des erreurs de raisonnement qu'a effectué l'agent pour pouvoir être cohérent avec son action. Par exemple M^a_{diag} correspond au fait que l'agent n'a pas porté son attention sur le fait qu'il soit en mode VS.

À partir de ces MCSes possibles, nous devons maintenant introduire l'*inertie* et trouver les différentes ignorances relatives à celle-ci. C'est pourquoi nous allons introduire l'*inertie* dans la troisième étape de notre algorithme que nous abordons dans la prochaine section.

VI.5.3 Ignorances liées à l'extrapolation

Nous avons dans les précédentes étapes trouvées une solution pour toutes les incohérences qui ne sont pas liées à l'*inertie* des croyances. Nous allons donc dans cette troisième étape introduire cette *inertie*. Nous allons dans un premier temps rechercher les ignorances qui sont dues au fait que l'agent met à jour ses croyances face à un monde qui évolue, par exemple l'agent croit que l'alarme ne sonne pas au pas de temps 0 puis sonne au pas de temps 1. C'est le problème de l'*extrapolation* de croyances comme nous avons vu sous-section V.3.1 qui s'intéresse aux changements dans les croyances au vu des observations de l'agent (et non de l'action effectuée qui correspond à la *mise à jour*).

Par conséquent, pour la troisième étape, nous allons retirer l'action et les désirs du système afin de ne pas considérer les croyances dues aux effets de l'action et les incohérences liés aux désirs mais uniquement des observations. De plus, nous allons introduire tous les littéraux d'*inertie* $\mathfrak{K}_t$ (voir section VI.3) pour ne considérer que les incohérences entre l'*inertie* des croyances et les croyances qui découlent des observations. Enfin, comme nous cherchons uniquement des mises à jour, nous allons autoriser les corrections seulement sur les littéraux d'*inertie* de valeur $\mathfrak{K}_t^{(val)}$ en protégeant tout l'état de croyance et les *inerties* sur les connaissances $\mathfrak{K}_t^{(known)}$. Nous notons M_{ext} un MCS possible de cette troisième étape et B_t^{ext} l'état de croyances résultant de M_{ext}. Le calcul des ignorances relatif à l'*extrapolation* correspond alors

à :

$$\begin{array}{ll} & B_t^{ext} = \Phi \setminus M_{ext} \\ \text{avec} & M_{ext} \in \mathfrak{M}\big(\Phi, screened\big) \\ & \Phi = \{B_t^{diag} \setminus \{a_t \cup \mathcal{D}\} \cup \mathfrak{K}_t\} \\ & screened = \{B_t^{diag} \cup \mathfrak{K}_t^{(known)}\} \end{array} \tag{VI.3}$$

Reprenons l'exemple du mont Sainte-Odile et considérons que nous souhaitons calculer la troisième étape à partir de l'état B_1^{diag} qui résulte du MCS $M_{diag}^a = \{\mathrm{onVS}_1\}$
de l'étape précédente (voir sous-section VI.5.2). Si nous recherchons B_t^{ext} avec le système $\Phi = \{B_t^{diag} \setminus a_t \cup \mathfrak{K}_t^{(val)}\}$ nous nous retrouvons avec une incohérence. En effet, l'agent peut observer qu'il n'est pas en bonne position au temps 1 alors qu'il croyait qu'il l'était au temps 0. Il y a donc un changement dans le monde que l'agent doit prendre en compte. Nous trouvons alors comme possibilité pour M_{ext} :

$$M_{ext}^a \quad = \quad \{keep_1^{(val)}(\mathrm{goodXaxis}), keep_1^{(val)}(\mathrm{goodYaxis})\}$$

Ici, il n'existe qu'une seule possibilité du fait que nous autorisons seulement les corrections sur les littéraux d'*inertie* pour considérer les croyances qui doivent être mises à jour. Dans cet exemple, seules les croyances sur la position de l'appareil doivent être mises à jour.

Nous avons ici pris en compte les changements attendus dans les croyances face à un changement dans le monde. Nous devons maintenant prendre en compte les incohérences entre l'action effectuée et l'*inertie* des croyances avec le problème de la *mise à jour* et de la *distorsion* que nous abordons dans la section suivante.

VI.5.4 Ignorances liées à la mise à jour et à la distorsion

Dans cette quatrième étape, nous nous intéressons aux incohérences entre l'action et l'*inertie* des croyances avec le problème de la *distorsion* et de la *mise à jour*. L'idéal aurait été de suivre là encore le principe par itération et d'avoir une étape pour la *mise à jour* et une étape pour la *distorsion*. Malheureusement, nous avons vu section VI.4 que calculer les solutions de la *distorsion* avec notre opérateur $\mathfrak{M}$ revenait à calculer un sur-ensemble de solutions à la *mise à jour*. De plus, nous ne pouvons pas avec notre opérateur $\mathfrak{M}$ calculer uniquement les solutions de la *mise à jour* sans trouver des solutions relatives à la *distorsion*. De ce fait, nous ne pourrons distinguer ces deux mécanismes qu'après avoir calculé l'ensemble des solutions. Cette distinction sera donc à la charge du modèle d' *évaluation* qui devra déterminer si ces ignorances sont une *mise à jour* ou une *distorsion*.

Dans les précédentes étapes, nous avons pu prendre en compte toutes les incohérences entre les observations et l'*inertie* des croyances. La quatrième étape consiste simplement à ajouter à nouveau l'action et les désirs dans l'état de croyance

pour déterminer les incohérences entre l'*inertie* et l'action. Nous notons M_{dist} un MCS possible de cette dernière étape et B_t l'état de croyances résultant de M_{dist}. Le calcul des ignorances relatif à la *distorsion* et à la *mise à jour* correspond alors à :

$$\begin{array}{ll} & B_t = \Phi \setminus M_{dist} \\ \text{avec} & M_{dist} \in \mathfrak{M}\big(\Phi, screened\big) \\ & \Phi = \{B_t^{ext} \cup \{a_t \cup \mathcal{D}\}\} \\ & screened = \{a_t, R^C\} \end{array} \tag{VI.4}$$

Reprenons l'exemple du mont Sainte-Odile et considérons que nous souhaitons calculer la quatrième étape à partir de l'état B_1^{ext} qui résulte du MCS $M_{ext} = \{keep_1^{(val)}(\mathrm{goodXaxis}), keep_1^{(val)}(\mathrm{goodYaxis})\}$ de l'étape précédente (voir sous-section VI.5.3). Si nous recherchons B_t avec le système $\Phi = \{B_t^{ext} \cup a_t\}$ nous nous retrouvons avec une incohérence. En effet, du fait des littéraux d'*inertie*, nous avons $\mathrm{onVS}_0 \longleftrightarrow \mathrm{onVS}_1$ avec onVS_0 à vrai et de la même manière que sous-section VI.5.2 l'agent peut conclure par les règles R^d et R^h que son désir d'être dans la bonne trajectoire n'est pas satisfait. Nous avons comme possibilité pour M_{dist} :

$$\begin{array}{lcl} M_{dist}^a & = & \{keep_1^{(val)}(\mathrm{onVS})\} \\ M_{dist}^b & = & \{keep_1^{(known)}(\mathrm{onVS})\} \\ M_{dist}^c & = & \{\mathrm{onVS}_0\} \\ M_{dist}^d & = & \{\mathcal{D}(\mathrm{goodYaxis}_2)\} \\ M_{dist}^e & = & \{R^d(1)\} \\ M_{dist}^f & = & \{R^h(1)\} \end{array}$$

Nous retrouvons par exemple ici une *distorsion* avec le MCS M_{dist}^a car l'agent change la valeur de sa croyance sur le mode VS et croit donc par la règle R^a qu'il est en mode FPA et non en mode VS tout en ne portant pas attention à l'information qui lui indique qu'il est en mode VS comme nous avons vu sous-section VI.5.2 avec le MCS M_{diag}^a.

VI.5.5 Synthèse de l'approche

Afin de faciliter la distinction des différentes natures d'ignorances possibles pour un état de croyances d'un agent, nous proposons un algorithme de diagnostic par itération. Le principe est de calculer un MCS pour chaque sous problème d'incohérence (*i.e révision, consistency-based diagnosis, extrapolation*, etc) qui ont pour solutions des ignorances de natures différentes (*i.e* préférence, erreur de raisonnement, mise à jour, etc). Pour chaque correction successive afin d'arriver à un état de croyance B_t à partir de B_{t-1}, il existe plusieurs solutions possibles. Nous avons alors une structure en arbre comme illustré figure VI.1 que nous appelons *arbre des corrections*. Un chemin de corrections x permettant de passer de B_{t-1} à B_t est donc une suite de MCSes $M_{rev}^x, M_{diag}^x, M_{ext}^x, M_{dist}^x$. L'ensemble des ignorances sur le chemin x, noté $\mathcal{I}_t^x$, correspond ainsi à l'union des MCSes sur de chemin dans

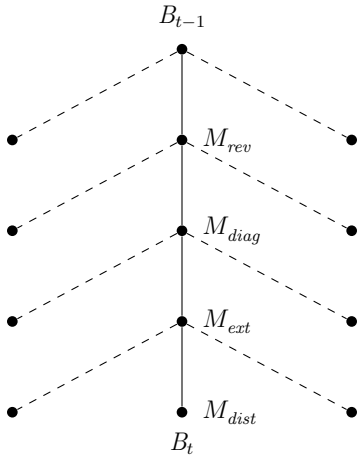


Figure VI.1 – Arbre des corrections de l'algorithme de diagnostic pour un état B_t

l'arbre des corrections : $\mathcal{I}_t^x = M_{rev}^x \cup M_{diag}^x \cup M_{ext}^x \cup M_{dist}^x$. Nous notons $\mathfrak{D}(\Phi)$ l'ensemble des chemins retourné par notre algorithme de diagnostic par itération sur le système Φ. Un état de croyances possible B_t^x correspond alors à :

$$\begin{array}{ll} & B_t^x = \Phi \setminus \mathcal{I}_t^x \\ \text{avec} & \Phi = \{B_{t-1} \cup \mathcal{D} \cup Obs_t \cup \{a_t\} \cup \mathcal{R} \cup \mathfrak{K}_t\} \\ & \mathcal{I}_t^x = \bigcup_{M \in path} M,\ path \in \mathcal{D}(\Phi) \end{array} \tag{VI.5}$$

Il y a autant de B_t^x que de chemins x possibles.

Le fait d'avoir des branchements dans notre algorithme de diagnostic d'un état de croyances pose la question de ce que représentent concrètement ces branchements et ces différentes corrections possibles. Nous allons voir dans la prochaine section que nous les interprétons comme des choix que l'agent peut effectuer.

VI.6 Définition du choix

Notre algorithme par itération nous permet de calculer un ensemble d'ignorances pour chaque sous problème et permet d'avoir une structure en arbre pour passer de l'état de croyance précédent au suivant. Cette structure en arbre nous permet de déterminer les différentes alternatives d'ignorances que l'agent pouvait effectuer. En effet, en calculant l'étape de révision M_{rev} chaque solution possible est une alternative de révision possible pour l'agent. Prenons un exemple :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\text{fireAlarm}_1, \neg\,\text{smoke}_1, \neg\,\text{smell}_1\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \text{fireAlarm}_t \rightarrow \text{fire}_t \\ R^b & \equiv & \neg\,\text{smoke}_t \rightarrow \neg\,\text{fire}_t \\ R^c & \equiv & \neg\,\text{smell}_t \rightarrow \neg\,\text{fire}_t \end{array} \right\} \end{array}$$

Ici l'agent observe une alarme incendie, mais pas de fumée ni d'odeur de fumée. Nous trouvons comme MCS pour l'étape de calcul des révisions possibles au pas de temps 1 :

$$\begin{array}{rcl} M^a_{rev} & \equiv & \{\mathrm{fireAlarm}_1\} \\ M^b_{rev} & \equiv & \{R^a_1\} \\ M^c_{rev} & \equiv & \{\neg\,\mathrm{smoke}_1, \neg\,\mathrm{smell}_1\} \\ M^d_{rev} & \equiv & \{R^b_1, \neg\,\mathrm{smell}_1\} \\ M^e_{rev} & \equiv & \{\neg\,\mathrm{smoke}_1, R^c_1\} \\ M^f_{rev} & \equiv & \{R^b_1, R^c_1\} \end{array}$$

L'agent a ici comme alternative de révision pour le MCS M^a_{rev} : M^b_{rev} ou M^c_{rev} ou M^d_{rev} ou M^e_{rev} ou M^f_{rev}. Chaque alternative est composée de différents choix d'ignorances. C'est-à-dire qu'une autre ignorance ψ a été choisie à la place de l'ignorance φ dans une alternative à un MCS. Par exemple, un autre choix possible pour l'ignorance de $\mathrm{fireAlarm}_1$ et l'ignorance de $\neg\,\mathrm{smoke}_1$ et de $\neg\,\mathrm{smell}_1$qui correspond à l'alternative M^c_{rev} : l'agent avait le choix, au lieu de ne pas considérer l'alarme incendie, choisir de ne pas considérer qu'il n'y a pas de fumée ni qu'il n'y a pas d'odeur.

Le même raisonnement peut être utilisé pour chaque étape de l'algorithme, les solutions possibles de l'étape M_{diag} représentent des choix d'ignorances pour une action cohérente, puis M_{ext} des choix d'extrapolation et M_{dist} des choix de distorsion. Nous allons donc maintenant définir formellement la notion de choix.

VI.6.1 Une recherche de Minimal Unsatisfaisable Sets

Un choix d'ignorance pour une proposition φ est finalement toute autre proposition logique ψ qui est incohérent avec φ. En effet, l'agent peut choisir d'ignorer ψ ou φ pour qu'il n'y ait plus d'incohérence entre ψ et φ. Par exemple, l'agent pouvait choisir l'ignorance $\neg\,\mathrm{smoke}_1$ à la place de $\mathrm{fireAlarm}_1$ dans l'exemple précédent, car les deux littéraux sont incohérents par la règle R^a. Nous recherchons donc les sous-ensembles minimaux incohérents dans les croyances de l'agent pour déterminer les différents choix d'ignorances que celui-ci aurait pu faire. Ces sous ensembles minimaux incohérents sont connus dans la littérature sous le nom de *Minimal Unsatisfaisable Sets* (MUSes) [Birnbaum et al., 2003]. En effet, les MUSes sont des sous ensembles minimaux incohérents de clauses logiques. C'est-à-dire qu'il suffit qu'une clause de chaque MUS soit retirée pour que le système devienne cohérent. Formellement :

$$MUSes(\Phi) \equiv \left\{ U \;\middle|\; \begin{array}{l} U \subseteq \Phi \\ U \vdash \bot \\ \forall U' \subset U, U' \nvdash \bot \end{array} \right\}$$

Si nous reprenons l'exemple de l'alarme incendie dans la section précédente, nous trouvons comme $MUSes$:

$$\begin{array}{rcl} U_1^a & \equiv & \{\mathrm{fireAlarm}_1, R_1^a, \neg\,\mathrm{smoke}_1, R_1^b\} \\ U_1^b & \equiv & \{\mathrm{fireAlarm}_1, R_1^a, \neg\,\mathrm{smell}_1, R_1^c\} \end{array}$$

Nous retrouvons donc bien le fait avec le U_1^a que $\mathrm{fireAlarm}_1$ est incohérent avec la règle R_1^a, $\neg\,\mathrm{smoke}_1$ et la règle R_1^b. De plus, nous retrouvons par le U_1^b que $\mathrm{fireAlarm}_1$ est aussi incohérent avec la règle R_1^c et l'odeur smell_1. Les MUSes nous donne donc l'information des autres choix possibles pour une proposition φ en déterminant les MUSes auquel φ appartient. Or pour calculer ces MUSes, [Birnbaum et al., 2003] ont montré une propriété intéressante avec les MCSes.

VI.6.2 Les hitting sets

En effet, les auteurs ont montré qu'il est possible de passer des MCSes aux MUSes et inversement en passant par les *hitting sets* : sachant un ensemble de contrainte insatisfaisable Φ

- Un sous ensemble M de Φ est un MCS de Φ si et seulement si M est un *hitting set* irréductible des MUSes de Φ
- Un sous ensemble Φ de Φ est un MUS de Φ si et seulement si U est un *hitting set* irréductible des MCSes de Φ.

Un ensemble H est un *hitting set* de l'ensemble d'ensemble $\mathcal{A}$, que nous notons $H\,hits\mathcal{A}$, si et seulement si chaque ensemble $A \in \mathcal{A}$ contient au moins un élément de H :

$$H\,hits\mathcal{A} \equiv (\forall A \in \mathcal{A},\ H \cap A \neq \emptyset)$$

Par exemple sur l'ensemble des MUSes $\{U_1^a, U_1^b\}$, deux *hitting sets* possible sont : $\{\mathrm{fireAlarm}_1\}$ et $\{\mathrm{fireAlarm}_1, \mathrm{smell}_1\}$

Un *hitting set* est dit *irréductible* si et seulement si aucune proposition ne peut être retiré sans que H ne soit plus un *hitting set* :

$$IHSs(\mathcal{A}) \equiv \left\{ H \;\middle|\; \begin{array}{l} H\,hits\mathcal{A} \\ \forall H' \subset H,\ \neg(H'\,hits\mathcal{A}) \end{array} \right\}$$

Dans notre exemple d'*hitting sets* $\{\mathrm{fireAlarm}_1, \mathrm{smell}_1\}$ n'est pas *irréductible* car en retirant smell_1, $\{\mathrm{fireAlarm}_1\}$ est toujours un *hitting set*.

Application Pour trouver les choix possibles pour une proposition φ dans un MCS M, nous devons dans un premier temps déterminer les MUSes en calculant les *hitting sets* sur les autres MCSes possibles que M (les alternatives de M). Un MCS résultant d'une étape de calcul (*i.e révision, extrapolation, distorsion*, etc), les autres MCSes possibles sont donc les voisins de M dans l'arbre des corrections de l'algorithme de diagnostic (voir figure VI.1). Si l'ensemble M résulte du calcul de l'étape de *révision de croyance* alors les autres MCSes possibles sont les voisins de M à l'étape de calcul de la *révision de croyance*. Nous notons la fonction $\mathcal{N}^{step}(M)$

qui retourne les voisins de M dans l'*arbre des corrections*. Nous définissons la fonction $\mathfrak{U}$ qui calcule les MUSes d'un MCS M par :

$$\mathfrak{U}(M) \equiv \left\{ U \;\middle|\; U \in \mathit{IHSs}(\mathcal{N}^{step}(M) \cup M_t) \right\}$$

À partir de là, nous devons déterminer les choix possibles sur les MUSes contenant φ car ceux-ci contiennent les propositions incohérentes avec φ. Si plusieurs MUSes contiennent φ, cela veut dire qu'il existe plusieurs propositions indépendantes logiquement qui sont incohérentes avec φ. Il faut donc piocher une proposition dans chaque MUS où φ est présent pour déterminer un autre choix de correction. Or cela correspond à la définition des *hitting set* irréductible. Nous allons donc cacluler à nouveau un *hitting set* irréductible sur les MUSes contenant φ pour déterminer les choix de φ. Nous notons alors la fonction de choix $choice$ qui prend en paramètres une proposition φ et un MCS M tel que $\varphi \in M$:

$$choice(\varphi, M) \equiv \left\{ \Psi \;\middle|\; \begin{array}{l} \Psi \in \mathit{IHSs}\left(\left\{ U \;\middle|\; \begin{array}{l} U \in \mathfrak{U}(M) \\ \varphi \in U \end{array} \right\} \right) \\ \varphi \notin \Psi \end{array} \right\}$$

$\varphi \notin \Psi$ permet de ne garder que les choix ne contenant pas φ.

Si nous reprenons les MCSes trouvés pour l'exemple de l'alarme à incendie, nous trouvons par exemple :

$$choice(\neg \mathrm{smoke}_1, M^c_{rev}) \quad \equiv \quad \{\{\mathrm{fireAlarm}_1\}, \{R^a_1\}, \{R^b_1\}\}$$

Ici la fonction *choice* nous dit que dans le MCS M^c_{rev}, le littéral $\neg \mathrm{smoke}_1$ peut être remplacé par la règle R^a_1, R^b_1 ou le littéral $\mathrm{fireAlarm}_1$ pour retrouver un autre MCS possible. Autrement dit, ignorer l'observation qu'il n'y a pas de fumée peut être mis en regard de trois autres choix :

— ignorer la signification de l'alarme (R^a) ;
— ignorer le fait que sans fumée, il n'y a pas de feux (R^b) ;
— ignorer l'alarme ($\mathrm{fireAlarm}_1$).

VI.6.3 Définition de l'ensemble alternatif d'ignorances

La fonction *choice* permet de déterminer quelles ignorances peuvent être remplacées par une autre, mais ne nous indique pas quel état de croyances résulte de ce choix d'ignorance : si l'agent a choisi de réviser ses croyances en ignorant l'alarme à la place de l'observation de la non précense de fumée, quel est l'ensemble d'ignorances possible le plus proche où ce choix de révision a été effectué ?

Pour trouver cet ensemble d'ignorances le plus proche qui résulte du choix, il ne suffit pas de remplacer l'ignorance φ par son choix ψ dans un ensemble d'ignorance $\mathcal{I}^x_t$. Considérons l'exemple précédent de l'alarme incendie. L'exemple ne contient qu'une incohérence liée à la révision de croyance, par conséquent seul l'étape du

calcul de M_{rev} contiendra des ignorances. Nous notons alors les ignorances possibles $\mathcal{I}_1^x$ pour le pas de temps 1 :

$$\begin{array}{lcl} \mathcal{I}_1^a & \equiv & \{\mathrm{fireAlarm}_1\} \\ \mathcal{I}_1^b & \equiv & \{R_1^a\} \\ \mathcal{I}_1^c & \equiv & \{\neg\,\mathrm{smoke}_1, \neg\,\mathrm{smell}_1\} \\ \mathcal{I}_1^d & \equiv & \{R_1^b, \neg\,\mathrm{smell}_1\} \\ \mathcal{I}_1^e & \equiv & \{\neg\,\mathrm{smoke}_1, R_1^c\} \\ \mathcal{I}_1^f & \equiv & \{R_1^b, R_1^c\} \end{array}$$

si nous remplaçons $\neg\,\mathrm{smoke}_1$ par le choix $\mathrm{fireAlarm}_1$ dans l'ignorance $\mathcal{I}_1^c$ nous trouvons les ignorances : $\{\mathrm{fireAlarm}_1, \neg\,\mathrm{smell}_1\}$ qui ne correspond pas à un ensemble d'ignorances possible, car il n'existe pas un autre ensemble d'ignorances $\mathcal{I}_1^x$ où tous ces littéraux sont ignorés. C'est pourquoi nous définissons un ensemble alternatif d'ignorances $\mathcal{I}_t^{alt}$ qui résulte d'un choix ψ pour $\varphi \in \mathcal{I}_t^x$ est alors :

— un ensemble d'ignorances retourné par notre algorithme de diagnostic (*i.e* dans l'arbre $\mathfrak{T}$) où ψ est ignoré ;
— un ensemble d'ignorances ayant un maximum d'ignorances en commun avec $\mathcal{I}_t^x$ (*i.e* le plus proche).

Formellement, nous notons la fonction alt qui retourne l'ensemble d'ignorances alternatif selon un choix Ψ pour un ensemble $\mathcal{I}_t$ d'ignorances :

$$alt(\Psi, \mathcal{I}_t^x) \quad = \quad \mathcal{I}_t^{alt} \left| \begin{array}{l} \mathcal{I}_t^{alt} \in \mathcal{N}(\mathcal{I}_t^x) \\ \Psi \in \mathcal{I}_t^{alt} \\ \nexists \mathcal{I}_t^y \in \mathcal{N}(\mathcal{I}_t^x),\, \Psi \in \mathcal{I}_t^y \wedge (|\mathcal{I}_t^y \cap \mathcal{I}_t^x| > |\mathcal{I}_t^{alt} \cap \mathcal{I}_t^x|) \end{array} \right.$$

où $\mathcal{N}(\mathcal{I}_t^x)$ est la fonction retournant les voisins des ignorances $\mathcal{I}_t^x$ dans l'arbre de diagnostic $\mathfrak{T}$. Dans notre exemple, nous recherchons l'ensemble alternatif d'ignorances où $\mathrm{fireAlarm}_1$ est ignoré à la place de $\neg\,\mathrm{smoke}_1$: c'est-à-dire $alt(\mathrm{fireAlarm}_1, \mathcal{I}_1^c)$. Ici seul $\mathcal{I}_1^a$ contient le choix $\mathrm{fireAlarm}_1$, il n'existe donc pas d'autre ensemble ayant plus d'ignorances en commun avec $\mathcal{I}_1^c$. Par conséquent, $\mathcal{I}_1^a$ est un ensemble alternatif pour $\mathcal{I}_1^c$ pour le choix de $\mathrm{fireAlarm}_1$: $alt(\mathrm{fireAlarm}_1, \mathcal{I}_1^c) = \mathcal{I}_1^a$.

Chaque étape de l'algorithme est donc le calcul d'un ensemble de choix d'ignorances minimales. Ces différents choix d'ignorances s'accumulent pour former un ensemble d'ignorance complet qui permet d'atteindre un état de croyances pour l'agent permettant d'expliquer son comportement à un instant t. L'accumulation de choix pour former un état de croyances pose la question de la complétude de notre algorithme par itération : Est-ce que nous ne passons pas à côté de correction minimale et par conséquent à des explications potentielles ? La prochaine section abordera la correction et complétude de l'algorithme et montrera formellement que celui-ci est complet et correct.

VI.7 Correction et complétude de l'algorithme

Notre algorithme par itération (que nous notons algorithme 2) nous permet de distinguer à chaque étape les ignorances de natures différentes relevant d'un problème d'incohérence différent (*i.e* révision, consistency-based diagnosis, extrapolation, mise à jour, distorsion). L'approche précédente pour calculer les ignorances (que nous notons algorithme 1), s'effectuait sur une unique étape où toutes les propositions étaient introduites pour calculer les MCSes en une seule fois. Ces ignorances étaient donc minimales. Les deux approches sont donc différentes, mais cherchent à calculer la même chose : les ignorances possibles de l'agent. Nous devons donc comparer ces deux approches en vérifiant notamment la correction et complétude de l'algorithme 2 :

(Correction) L'algorithme 1 retournant un ensemble minimal d'ignorance pour rendre les croyances cohérentes, nous devons vérifier que l'algorithme 2 ignore au moins la même chose que l'algorithme 1 pour rendre les croyances cohérentes.

(Complétude) L'algorithme 1 retournant uniquement des solutions minimales, nous devons vérifier que les solutions minimales sont aussi retournées par l'algorithme 2 afin de confirmer que l'algorithme 2 ne passe pas à côté d'une solution d'ignorance possible.

Nous allons dans les prochaines sections justifier formellement que l'algorithme par itération est correct et complet. Nous discuterons ensuite des différentes solutions retrouvées par l'algorithme par itération par rapport à l'approche précédente.

VI.7.1 Correction

Nous allons abstraire l'algorithme sans itération (que nous notons algorithme 1) et avec itération (que nous notons algorithme 2) pour faciliter les preuves de correction et de complétude. Considérons que l'algorithme 1 consiste à corriger l'ensemble B pour le rendre cohérent. Finalement, l'algorithme 2 correspond à cinq itérations où une itération consiste à :

1. ajouter des propositions de B dans un sous ensemble A de B ;
2. corriger A.

Ces itérations sont répétées jusqu'à ce que toutes les propositions manquantes au sous ensemble A de la première itération par rapport à B ont été ajoutées. Ainsi, si nous montrons la correction et la complétude sur deux itérations, cela est équivalent à la montrer sur cinq itérations du fait que nous ne faisons qu'ajouter des propositions dans un sous ensemble pour arriver à B : c'est un algorithme récursif.

Ainsi, l'algorithme 2 est abstrait sur deux pas de temps et consiste à calculer un MCS M_1 sur un ensemble A puis de calculer un MCS M_2 sur l'ensemble $B \setminus M_1$ où $A \subseteq B$. L'ignorance qui résulte de ces deux pas de temps est alors l'union des deux MCses : $M_1 \cup M_2$. Pour le cas de l'algorithme 1, celui-ci consiste à calculer un MCS M sur l'ensemble B. L'ignorance qui résulte de cet algorithme est alors équivalent à M.

Déterminer si l'algorithme 2 est correct revient à déterminer si les ignorances trouvées par l'algorithme 1 est un sous ensemble des ignorances trouvées par l'algorithme 2, c'est-à-dire que l'algorithme 2 ignore au moins les mêmes propositions que l'algorithme 1 :

$$(T4) \quad \begin{array}{l} \text{Si } A \subseteq B \wedge M_1 \in \mathfrak{M}\big(A, \emptyset\big) \wedge M_2 \in \mathfrak{M}\big(B \setminus M_1, \emptyset\big) \\ \text{alors } \exists M \in \mathfrak{M}\big(B, \emptyset\big),\ M \subseteq M_1 \cup M_2 \end{array}$$

Nous pouvons affirmer que :

(a) $(B \setminus (M_1 \cup M_2)) \subseteq B$ du fait que retirer des ensembles à B ne peut donner qu'un sous ensemble de B.

(b) $(B \setminus (M_1 \cup M_2)) \nvdash \bot$ car par construction, M_2 rend $B \setminus M_1$ cohérent (c'est un MCS) donc a fortiori $B \setminus (M_1 \cup M_2)$ ne peut être que cohérent.

(c) $B \setminus M$ est un *remainder* de B pour faux, c'est-à-dire un sous ensemble maximal qui n'implique pas faux. En effet, M étant un MCS de B par l'équivalence entre MCS et *remainder* démontré section V.4, nous pouvons conclure que M est un *remainder*.

Nous savons donc par (c) que M est minimale afin que $B \setminus M$ soit un sous-ensemble maximal de B qui n'implique pas faux. De plus, nous savons par (a) et (b) que $B \setminus (M_1 \cup M_2)$ est un sous ensemble de B qui n'implique pas faux. Par conséquent, $M_1 \cup M_2$ ne peuvent être qu'un sur-ensemble d'un MCS M de B. En effet, du fait que M est minimale et rend B cohérent, ainsi que du fait que $M_1 \cup M_2$ rend aussi B cohérent, l'ensemble $M_1 \cup M_2$ ne peut être qu'un sur-ensemble de M. Dans le cas contraire, cela voudrait dire que M ne serait pas minimale, ce qui est une contradiction.

Nous pouvons donc bien en conclure que $\exists M \in \mathfrak{M}\big(B, \emptyset\big),\ M \subseteq M_1 \cup M_2$ et qu'ainsi le théorème $(T4)$ de correction est vrai.

Nous allons maintenant dans la prochaine section déterminer si l'algorithme 2 est complet.

VI.7.2 Complétude

Le théorème $(T4)$ sous-section VI.7.1 nous indique que les solutions trouvées par l'algorithme 2 contiennent au moins les ignorances trouvées par l'algorithme 1. De plus, il nous indique que les solutions trouvées par itération ne sont pas forcément minimales, c'est-à-dire que par itération, il est possible de trouver des solutions où l'agent ignore plus de chose que sans itération. Déterminer si l'algorithme 2 est complet revient alors à déterminer si celui-ci retourne aussi des

solutions minimales d'ignorances et ainsi retrouve les mêmes solutions que l'algorithme 1. Plus formellement, pour toute solution minimale M, est-ce qu'il existe M_1 et M_2 obtenus par l'algorithme 2 tels que $M_1 \cup M_2 = M$:

$$(T5) \quad \begin{array}{l} \text{Si } A \subseteq B \wedge M \in \mathfrak{M}\big(B, \emptyset\big) \\ \text{alors } \exists M_1, M_2,\ M_1 \in \mathfrak{M}\big(A, \emptyset\big) \wedge M_2 \in \mathfrak{M}\big(B \setminus M_1, \emptyset\big) \\ \qquad \wedge M = M_1 \cup M_2 \end{array}$$

Pour commencer, notons que si $B \setminus M$ est cohérent et $A \subseteq B$, alors non seulement $A \setminus M$ est cohérent (il y a moins de propositions) mais de plus il existe un $M' \subseteq M$ tel que $A \setminus M'$ est cohérent (certaines ignorances de M ne sont pas présentes dans A donc nous pouvons les enlever). Nous pouvons affirmer à partir de là que :

(a) $\forall A \subseteq B, \exists M_1 \in \mathfrak{M}\big(A, \emptyset\big)$ avec $M_1 \subseteq M$.

(b) $\forall M_1, \exists M_2 \in \mathfrak{M}\big(B \setminus M_1, \emptyset\big)$ avec $M2 \subseteq M$.

Nous en déduisons alors :

(c) $M_1 \cup M_2 \subseteq M$ par (a) et (b).

(d) d'après $T4$ (voir sous-section VI.7.1), $\exists M' \in \mathfrak{M}\big(B, \emptyset\big)$ tq $M_1 \cup M_2 = M'$.

(e) Nous avons donc $M' = M_1 \cup M_2$ et $M' \subseteq M$ mais comme M et M' sont tous le deux des MCS, par définition de la minimalité, $M' = M$.

L'algorithme 2 est donc complet dans le sens qu'il retourne comme l'algorithme 1 les ignorances minimales qui permettent que les croyances de l'agent soient cohérentes. Toutefois, bien que l'algorithme 2 retrouve les mêmes solutions, les théorèmes $(T4)$ et $(T5)$ nous indique aussi qu'il existe aussi des solutions qui ne sont pas minimales. En effet, $T5$ indique que toute solution minimale est une solution de l'algorithme 2. En revanche, le théorème $T4$ ne permet pas de conclure la réciproque, c'est-à-dire que pour toutes solutions de l'algorithme 2 ce n'est pas forcément une solution minimale car les solutions de l'algorithme 2 peuvent être des sur-ensembles d'une solution minimale. Nous allons discuter de cette dernière possibilité dans la prochaine section.

VI.7.3 Discussion

Par le théorème $(T4)$ et $(T5)$ nous savons que l'union des MCSes qui résulte de l'algorithme 2 peut être :

— un sur-ensemble d'une solution minimale trouvé par l'algorithme 1 $(T4)$;

— une solution minimale équivalente aux solutions trouvées par l'algorithme 1 $(T5)$.

D'un point de vue purement diagnostic en logique, le *consistency-based diagnosis* (section III.2) définit les meilleures solutions pour expliquer un comportement non attendu comme les solutions minimales qui permettent de retrouver la cohérence. Du fait que l'algorithme 2 retourne des solutions non minimales, nous pouvons nous poser la question de leurs pertinences pour expliquer le comportement de l'agent.

Nous pensons que, bien que l'algorithme 2 retourne des solutions où l'agent ignore plus de chose que nécessaire, ces solutions ne sont néanmoins pas inintéressantes. Pour illustrer notre propos, considérons un exemple simplifié du premier pas de temps de l'accident du vol Rio-Paris :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\mathrm{alarm}_1, \mathrm{acceleration}_1\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \mathrm{alarm}_t \rightarrow \mathrm{stall}_t \\ R^b & \equiv & \mathrm{acceleration}_t \rightarrow \neg\, \mathrm{stall}_t \\ R^c & \equiv & [\neg\, \mathrm{stall}_t]\, \mathrm{pull}_t \end{array} \right\} \\ a_1 & \equiv & \mathrm{pull}_1 \end{array}$$

Dans cet exemple où l'agent décide de tirer le manche de contrôle face à deux informations contradictoires, nous trouvons pour l'algorithme 1 les solutions suivantes :

$$\begin{array}{lcl} \mathcal{I}_1^a & \equiv & \{\mathrm{alarm}_1\} \\ \mathcal{I}_1^b & \equiv & \{R_1^a\} \\ \mathcal{I}_1^c & \equiv & \{R_1^c, R_1^b\} \\ \mathcal{I}_1^d & \equiv & \{R_1^c, \mathrm{acceleration}_1\} \end{array}$$

Pour l'algorithme 2, nous trouvons les solutions suivantes :

$$\begin{array}{lcl} \mathcal{I}_1^a & \equiv & \{\mathrm{alarm}_1\} \\ \mathcal{I}_1^b & \equiv & \{R_1^a\} \\ \mathcal{I}_1^c & \equiv & \{R_1^c, R_1^b\} \\ \mathcal{I}_1^d & \equiv & \{R_1^c, \mathrm{acceleration}_1\} \\ \mathcal{I}_1^e & \equiv & \{\mathrm{acceleration}_1, \mathrm{alarm}_1\} \\ \mathcal{I}_1^f & \equiv & \{\mathrm{acceleration}_1, R_1^a\} \\ \mathcal{I}_1^d & \equiv & \{R_1^b, R_1^a\} \\ \mathcal{I}_1^g & \equiv & \{R_1^b, \mathrm{alarm}_1\} \end{array}$$

Ainsi l'algorithme 2 retourne bien les mêmes solutions que l'algorithme 1 mais donne en plus des solutions non minimales ($\mathcal{I}_1^e$ à $\mathcal{I}_1^g$). Ces solutions explorent des révisions de croyance qui vont à contre-sens de la décision prise par l'agent. Par exemple, dans $\mathcal{I}_1^e$ et $\mathcal{I}_1^f$, l'agent ne prend pas en compte l'information de l'accélération alors qu'elle va dans le sens de sa décision. Autrement dit, il n'est pas strictement nécessaire d'ignorer l'accélération puisqu'elle ne contredit pas, in fine, la décision de l'agent. Mais elle provient de ce que l'agent a dû gérer une incohérence lors de la phase de *révision*. Ainsi, l'algorithme 1 n'explore pas ce genre de solution : la révision de croyance choisie par l'agent est toujours cohérente avec sa décision. Au contraire, l'algorithme 2 permet d'explorer des comportements plus complexes où par exemple l'agent préfère une information plutôt qu'une autre tout en ignorant des règles de raisonnement permettant d'utiliser correctement cette information. On peut considérer par exemple que $\mathcal{I}_1^f$ peut signifier que l'agent porte son attention sur l'alarme à la place de l'accélération, mais considère que l'alarme est défectueuse et qu'elle n'indique pas un décrochage.

L'algorithme 2 permet donc d'avoir une complexité d'état de croyances plus élevée tout en permettant une compréhension plus facile des ignorances. Or l'algorithme 1 comme nous avons noté sous-section V.5.5 retournait déjà de grand nombre de scénarios qui nécessitaient une fonction d'*évaluation* afin de déterminer uniquement les scénarios les plus plausibles. En générant encore plus de scénario avec l'algorithme 2, la nécessité d'une telle fonction est accentuée. C'est pourquoi nous aborderons dans le prochain chapitre le *modèle d'évaluation* afin de répondre à cette problématique.

VI.8 Conclusion

Dans ce chapitre, nous avons pu voir que l'*inertie* des croyances qui est essentiel dans un agent rationnel (voir sous-section III.1.5), introduit dans le contexte du diagnostic un nouveau problème à prendre en compte : *le problème de la distorsion du décor*. Ce problème consiste à prendre en compte des *distorsions* dans l'*inertie* des croyances, c'est-à-dire des changements qui ne sont pas dus à *une extrapolation* ou à une *mise à jour*, qui permettent d'expliquer une action incohérente de l'agent. Nous avons pu voir notamment que l'accident du mont Saint-Odile peut être expliqué par ces *distorsions* qui représentent dans le cas de celui-ci un oubli sur le mode de configuration de l'autopilote.

De plus, nous avons pu voir que nous prenons en compte l'inertie dans le modèle à travers deux littéraux d'*inertie* :

— $keep_t^{(val)}(\varphi)$ qui force la valeur de vérité de φ_t à être la même que φ_{t-1} ;

— $keep_t^{(known)}(\varphi)$ qui force la connaissance de φ_t à être la même que φ_{t-1}.

C'est l'ignorance de ces littéraux qui permet de représenter une *extrapolation*, *mise à jour* ou *distorsion*. En effet, nous avons pu voir dans ce chapitre que ces trois mécanismes correspondent à une *révision de croyance minimale* à la AGM et peuvent par conséquent être capturés par notre opérateur de diagnostic $\mathfrak{M}$ abordé chapitre V.

Du fait des différents mécanismes qui entraînent des ignorances pour retrouver la cohérence, ces ignorances sont de natures différentes. Une ignorance dû à une *révision de croyance* représente une préférence entre différentes croyances, l'*extrapolation* représente une mise à jour des croyances face à une évolution du monde, la *distorsion* une erreur dans l'*inertie* des croyances, etc. Afin de faciliter la distinction de ces différentes natures, nous avons proposé un algorithme de diagnostic par itération qui consiste à résoudre les incohérences de chaque problématique un à un. De ce fait, l'ensemble d'ignorances final permettant de rendre l'état de croyances B_t cohérent peut être décomposé en sous-ensembles représentant des ignorances de natures différentes. Il est à noter toutefois que notre opérateur ne permet pas de faire la distinction entre la *mise à jour* et la *distorsion*, il sera donc nécessaire de la faire après coup.

Enfin, nous avons pu prouver formellement que notre algorithme de diagnostic

par itération est correct et complet : nous ne passons pas à côté d'une explication minimale. Toutefois, l'algorithme par itération ressort plus de solutions que la version sans itération présentée chapitre V. Ainsi l'algorithme par itération permet de faciliter la compréhension et l'expressivité des ignorances, mais dans le même temps augmente le nombre d'états de croyances trouvé.

Nous avons donc un modèle d'explication composé d'un modèle $\mathcal{M}^l$ qui représente l'accident du point de vue de l'agent et retourne un ensemble de scénario $\mathfrak{T}$ qui représente un ensemble de séquences d'ignorances expliquant les actions de l'agent en prenant en compte des *distorsions* possibles dans l'*inertie* des croyances de l'agent. Toutefois, comme développé chapitre IV, ce modèle répond à la question du *comment* de l'action effectuée mais pas du *pourquoi*. Ainsi, ce modèle ne donne pas une plausibilité sur les états de croyance qui permettrait de considérer uniquement les scénarios les plus probables d'un point de vue cognitif. Nous allons donc dans le chapitre suivant définir le modèle d'*évaluation* en charge de donner un sens à ces ignorances, notamment à l'aide des biais cognitifs.

VII - Le modèle d'évaluation

Nous avons pu voir dans les chapitres précédents que le modèle d'*explication* à partir d'un modèle $\mathcal{M}^l$ permet d'obtenir un ensemble de scénario sous la forme d'un arbre d'ignorances $\mathfrak{T}$, c'est-à-dire un ensemble de séquences d'ignorances que l'agent a effectuées et qui explique ses actions. Nous avons vu chapitre IV que le modèle d'*explication* permet de répondre à la question du *comment* de l'action incohérente mais pas du *pourquoi*. C'est-à-dire que les ignorances n'ont pas pour l'instant un sens du point de vue cognitif mais correspondent seulement à des propositions que l'agent a ignorées pour retrouver la cohérence. Le but de ce modèle d'*évaluation* est donc de déterminer si une ignorance est plausible en recherchant une explication cognitive à cette ignorance.

Nous avons expliqué en sous-section IV.2.5, que pour atteindre cet objectif, nous aurons besoin d'établir trois définitions :

(1) la définition d'une taxonomie formelle des caractéristiques des biais cognitifs ;
(2) la définition des biais cognitifs basée sur cette taxonomie ;
(3) la définition d'une évaluation de la plausibilité d'un état de croyance en fonction des biais trouvés.

Nous allons dans ce chapitre développer ces 3 points afin de définir le modèle d'*évaluation*.

VII.1 Une taxonomie d'ignorances, de croyances et de choix

Nous avons pu voir chapitre II qu'il n'existe pas de consensus sur une taxonomie des biais cognitifs. Nous ne pouvons donc pas simplement formaliser une taxonomie existante car le choix d'une taxonomie plutôt qu'une autre serait purement subjectif du fait qu'elles reposent sur des caractéristiques floues ou imprécises qui par définition sont en opposition à un cadre formel. Nous pensons donc qu'il est préférable de développer notre propre taxonomie à partir du cadre formel que propose le modèle d'*explication*. En effet, le modèle d'*explication* retourne un arbre de scénario de diagnostic $\mathfrak{T}$ où chaque nœud correspond à des ignorances possibles

$\mathcal{I}_t$ à un pas de temps t (voir sous-section V.3.4). Un chemin dans cet arbre correspond à un *scénario* S_i. D'une ignorance $\mathcal{I}_t$ nous pouvons reconstruire un état de croyances de l'agent noté $B_t^{\mathcal{I}_t}$ qui correspond aux croyances que l'agent avait au pas de temps t dans ce scénario, en tenant compte des ignorances $\mathcal{I}_t$. Ce sont donc ces ignorances et ces états de croyances que nous devons caractériser formellement pour définir une taxonomie de biais cognitifs car ils correspondent dans notre cadre à l'expression des erreurs cognitives de l'agent.

Nous allons dans un premier temps définir formellement ce que l'on entend par caractéristiques puis nous verrons que ces caractéristiques peuvent être définies selon trois niveaux d'analyse : *intrinsèque, local* et *global*. Nous verrons que cette analyse repose sur trois principaux concepts : les ignorances, les états de croyances et les choix que nous définirons formellement. Enfin, nous verrons pour chaque niveau d'analyse la définition des différentes caractéristiques.

VII.1.1 Définition d'une caractéristique

Nous définissons une caractéristique comme une conjonction de conditions qui permet de différencier une ignorance φ d'autres ignorances ou un état de croyances B d'autres états de croyances. Ces caractéristiques ne permettent pas à elles seules de comprendre les raisons de l'ignorance ou de l'état de croyances, mais identifient des propriétés qui peuvent faire partie de cette explication. En d'autres termes, ce sont les briques de base pour construire la taxonomie de biais. Par exemple, une caractéristique peut être le fait que l'agent ignore une observation φ. Cela permet d'offrir un début d'explication pour l'ignorance φ, du fait que l'agent n'a pas porté son attention sur φ, mais n'offre à elle seule pas une explication cognitive de l'ignorance. Ces différentes briques seront ensuite combinées entre elles pour offrir une explication complète d'une ignorance φ.

Formellement, nous notons une caractéristique comme une fonction booléenne $\mathcal{F}$ qui prend un ensemble de paramètres Ψ et retourne vrai si Ψ est caractérisé par $\mathcal{F}$:

$$\mathcal{F} : \Psi \longmapsto \text{Bool}$$

La fonction $\mathcal{F}$ correspond à une conjonction de condition booléenne $\alpha_1 \wedge \cdots \wedge \alpha_n$ que nous représentons ainsi :

$$\mathcal{F}(\Psi) \Leftrightarrow \begin{cases} \alpha_1 \\ \vdots \\ \alpha_n \end{cases}$$

Pour définir une caractéristique, nous devons donc définir une fonction $\mathcal{F}$ à travers ses conditions $\{\alpha_1, \ldots, \alpha_n\}$ correspondantes.

Avant de définir un ensemble de fonctions $\mathcal{F}$ permettant de caractériser les ignorances, nous allons dans la prochaine section voir que les fonctions $\mathcal{F}$ peuvent être définies à partir de plusieurs granularités.

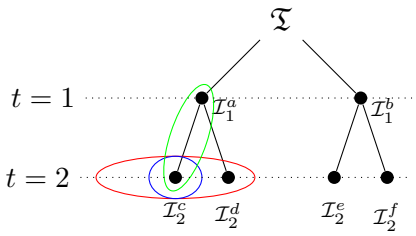


Figure VII.1 – Trois granularités

VII.1.2 Trois granularités d'analyse

Nous avons vu sous-section V.3.4 que l'ensemble des scénarios $\mathfrak{T}$ peut être vu comme une structure en arbre où chaque nœud est un ensemble d'ignorances $\mathcal{I}_t^x$ et chaque chemin un scénario. Considérons que nous cherchons à caractériser l'ignorance $\mathcal{I}_2^c$ sur la figure figure VII.1. Par cette structure en arbre, nous pouvons déterminer des caractéristiques sur trois granularités, comme nous pouvons le voir sur la figure VII.1 par les cercles de couleurs. :

1. En bleu, nous parlons de caractéristiques *intrinsèques*, c'est-à-dire des caractéristiques qui permettent de définir des propriétés propres aux ignorances de $\mathcal{I}_2^c$ et de l'état de croyances B_2^c. Par exemple, une explication possible pour avoir tiré le manche de contrôle dans l'accident du Rio-Paris est d'ignorer l'observation de l'alarme de décrochage, ce qui nous renseigne sur le fait que le pilote a fait une erreur d'inattention. Cette caractéristique se situe uniquement au niveau de l'état de croyances, sans tenir compte des autres états de croyances possibles ni des états précédents dans l'arbre.
2. En rouge, nous parlons de caractéristiques *locales*, c'est-à-dire des caractéristiques qui permettent de définir des propriétés sur $\mathcal{I}_2^c$ et B_2^c par rapport aux autres *choix* possibles (voir section VI.6) sur la même branche de scénario : dans notre exemple, le chemin d est une alternative au chemin c dans l'arbre des corrections. Il s'agit de comparer deux états de croyances situés au même niveau dans l'arbre et issus du même état précédent (parent). Par exemple, qu'est-ce qui peut expliquer le choix de préférer d'ignorer l'alarme de décrochage plutôt qu'une règle de raisonnement dans l'accident Rio-paris ?
3. En vert, nous parlons de caractéristiques *globales*, c'est-à-dire des caractéristiques qui permettent de définir des propriétés sur $\mathcal{I}_2^c$ et B_2^c par rapport aux ignorances et états précédents dans le scénario (*i.e* $\mathcal{I}_1^a$ et B_1^a). Dans le Rio-Paris toujours, le scénario privilégié par le BEA est que les pilotes ont cru être en survitesse. Cela correspond dans notre modèle à des ignorances successives qui permettent de croire à la survitesse de l'avion et de maintenir cette croyance au cours du temps.

Nous pouvons donc définir des caractéristiques propres aux ignorances et à un

état de croyance ou en comparaison à des *choix* possibles ou aux états de croyances et ignorances précédentes. Notre taxonomie va ainsi définir les caractéristiques à partir de trois concepts principaux : les ignorances, états de croyances et *choix*. Nous avons déjà donné une définition formelle des ignorances et des états de croyances chapitre V ainsi que du *choix* en chapitre VI.

Dans la suite de ce manuscrit nous garderons le même code couleur pour la définition des caractéristique. Les caractéristiques définissent par un label en bleu correspondent à des caractéristiques *intrinsèques*, en rouge à des caractéristiques *locales* et en vert *globales*.

VII.1.3 Caractéristiques intrinsèques

Pour définir les caractéristiques intrinsèques aux ignorances et aux états de croyances, nous nous plaçons sur une granularité d'analyse où nous avons accès aux ignorances $\mathcal{I}_t^x$ de l'agent à un pas de temps t, correspondant au choix du chemin de correction x entre B_{t-1} et B_t. De cet ensemble d'ignorances, nous pouvons en déduire l'état de croyances $B_t^{\mathcal{I}^x}$ (voir section VII.1). De plus, nous savons que l'ensemble d'ignorances $\mathcal{I}_t^x$ peut être découpé en sous-ensembles $\mathcal{I}_t^x = M_{rev}^x, M_{diag}^x, M_{ext}^x, M_{dist}^x$ où chaque sous ensemble représente des ignorances de nature différente (voir section VI.5). Enfin, nous savons que chaque littéral $\varphi_{t'}$ ignoré par $\mathcal{I}_t^x$ fait partie d'un ensemble du modèle $\mathcal{M}^l$ (*e.g* les règles, observations, désirs...).

De ce fait, si nous cherchons à caractériser une ignorance $\varphi_{t'} \in \mathcal{I}_t^x$ sur cette granularité, nous pouvons distinguer deux types de caractéristiques :

- Le type d'ignorance qui correspond à déterminer à quel ensemble de $\mathcal{M}^l$ l'ignorance φ appartient pour déterminer le processus cognitif en cause dans cette ignorance (*i.e* attention, mémoire, raisonnement, etc).
- La nature de l'ignorance qui correspond à déterminer à quel sous-ensemble de $\mathcal{I}_t^x$ l'ignorance $\varphi_{t'}$ appartient pour déterminer à quel problème d'incohérence l'ignorance est dû (*i.e* la révision, mise à jour, distorsion, etc).

Enfin, si nous cherchons à caractériser un état de croyances sur cette granularité, nous pouvons déterminer des caractéristiques liées aux désirs en fonction des désirs satisfaits ou non par l'état de croyances.

Nous allons définir dans les sections suivantes les caractéristiques liées aux types d'ignorance, à la nature de l'ignorance et aux désirs.

VII.1.3.a Type d'ignorance

En fonction de l'ensemble du modèle $\mathcal{M}^l$ auquel appartient l'ignorance $\varphi_{t'}$, elle représente une ignorance dû à un processus cognitif différent. Par conséquent, deux ignorances dans deux ensembles du modèle $\mathcal{M}^l$ différents représentent différents effets. Par exemple, ignorer une proposition $\varphi_{t'}$ qui est dans les observations veut dire que l'attention de l'agent n'était pas sur $\varphi_{t'}$ ce qui a pour effet que l'agent

ne prend pas en compte l'information $\varphi_{t'}$. Les types d'ignorance nous donnent donc des indices sur le mécanisme cognitif qui a entraîné l'ignorance, mais ne donne pas une explication complète de celle-ci. En effet, la non prise en compte de l'information peut-être dû à un choix entre deux informations contradictoires, une attention poussée vers une information négative, etc, en d'autres termes, d'autres caractéristiques qui doivent être définies.

Nous proposons de définir une caractéristique pour chaque type d'ignorance, c'est-à-dire pour chaque ensemble pouvant être ignoré dans les différente étapes de notre algorithme de diagnostic (voir section VI.5).

Non prise en compte d'une information À un instant t, l'agent a à sa disposition un ensemble d'observations. L'agent ayant des ressources cognitives limitées, son attention ne peut être omnisciente et par conséquent l'agent peut ne pas porter son attention sur toutes les informations. Une ignorance d'une observation vient alors traduire le fait que l'agent n'a pas finalement pris en compte l'information. Nous définissons la caractéristique *inObs* qui retourne vraie si $\varphi_{t'}$ est dans les observations du pas de temps t :

$$\mathbf{inObs}(\varphi_{t'}, t) \Leftrightarrow \varphi_{t'} \in Obs_t$$

Non-application d'un raisonnement L'agent ayant des ressources limitées, il est possible que celui-ci n'applique pas toutes les règles de raisonnement pour inférer toutes les croyances possibles. Nous capturons ce comportement si l'agent décide d'ignorer une règle de raisonnement, car cela traduit le fait que l'agent n'applique pas une règle et n'infère pas les informations qui en découlent. Nous définissons la caractéristique *isRule* qui retourne vraie si $\varphi_{t'}$ est une règle de raisonnement :

$$\mathbf{isRule}(\varphi_{t'}) \Leftrightarrow \varphi_{t'} \in \mathcal{R}$$

De plus, nous pouvons distinguer quatre cas particuliers pour les règles de raisonnement. En effet, une règle peut être une règle de précondition ou une règle d'effet de l'action effectuée et cela ne traduit pas la même chose. Ignorer les préconditions de l'action revient à une violation de règle. Par exemple, si un pilote doit continuer son trajet si et seulement si aucun danger ne se présente devant lui et que celui-ci décide d'ignorer cette règle, alors nous pouvons parler de violation de règle préétablie dans le raisonnement d'un pilote. Nous définissons alors la caractéristique *isCondition* qui retourne vrai si $\varphi_{t'}$ est une règle de précondition de l'action du pas de temps t :

$$\mathbf{isCondition}(\varphi_{t'}, t) \Leftrightarrow \varphi_{t'} \in prec(a_t)$$

avec $prec(a_t)$, l'ensemble des règles de préconditions de l'action a_t. Dans le cas de l'ignorance d'une règle d'effet d'action, cela traduit que l'agent croit que les effets de l'action effectuée ne sont pas ceux qui sont attendus. Par exemple, le pilote peut croire que pousser le manche de contrôle pendant un décrochage ne le fera pas sortir de la situation de décrochage. Nous définissons alors la caractéristique

isConsequence qui retourne vraie si $\varphi_{t'}$ est une règle d'effet de l'action du pas de temps t :

$$\textbf{isConsequence}(\varphi_{t'}, t) \Leftrightarrow \varphi_{t'} \in post(a_t)$$

avec $post(a_t)$ l'ensemble des règles de post-condition de l'action a_t.

Non prise en compte de l'inertie L'agent ayant des ressources limitées peut avoir des problèmes de mémoire et peut par exemple potentiellement oublier une information qu'il connaissait par le passé. Nous capturons cet oubli par l'ignorance d'une proposition d'*inertie* de connaissance. Nous notons alors la caractéristique $isKnownInertia$ qui retourne vraie si $\varphi_{t'}$ est une proposition de type $keep^{(known)}$ pour le pas de temps t :

$$\textbf{isKnownInertia}(\varphi_{t'}, t) \Leftrightarrow \varphi_{t'} \in \mathfrak{K}_t^{(known)}$$

Enfin, l'ignorance d'une proposition d'*inertie* de valeur peut correspondre à un rappel d'information erronée. Par exemple, l'agent croit que son pistolet n'est pas chargé alors qu'il croyait précédemment qu'il était chargé. Toutefois, l'ignorance d'une telle proposition d'*inertie* peut correspondre aussi à une *mise à jour* : l'agent croit que son pistolet n'est pas chargé car il a fait l'action de le décharger. Il est donc nécessaire de prendre en compte la caractéristique de la nature de l'ignorance pour faire la distinction de ces deux cas. Nous notons toutefois la caractéristique *isValueInertia* qui retourne vrai si $\varphi_{t'}$ est une proposition de type $keep^{(val)}$ pour le pas de temps t :

$$\textbf{isValueInertia}(\varphi_{t'}, t) \Leftrightarrow \varphi_{t'} \in \mathfrak{K}_t^{(val)}$$

Abandon de désir Si l'agent ignore une proposition $\varphi_{t'}$ qui est dans les désirs, cela veut dire que l'agent se retrouve dans un état de croyances où son désir est abandonné. Cela peut traduire alors un état où l'agent a des émotions négatives du fait qu'un désir n'est pas satisfait et peut être un indice de la prise de décision de l'agent. Par exemple, un agent qui décide de faire une action risquée pour sortir de l'état où son désir n'est pas satisfait. Nous notons alors la caractéristique *isDesire* qui retourne vrai si $\varphi_{t'}$ est dans les désirs de l'agent :

$$\textbf{isDesire}(\varphi_{t'}) \Leftrightarrow \varphi_{t'} \in \mathcal{D}$$

VII.1.3.b Nature de l'ignorance

En fonction du sous-ensemble de $\mathcal{I}_t$ auquel appartient $\varphi_{t'}$ alors $\varphi_{t'}$ n'est pas de même nature car l'ignorance permet de résoudre un problème d'incohérence différent comme nous avons vu section VI.4. L'ignorance peut être due à un problème de :

- Cohérence entre des croyances contradictoires qui est détectée pendant l'étape de *révision*. Nous parlons alors d'*ignorance de préférence*.
- Décision incohérente avec une action effectuée avec l'agent qui est détectée pendant l'étape de *diagnostic*. Nous parlons d'*ignorance de décision*.

— Décision incohérente avec avec les croyances de l'agent qui est détectée pendant l'étape de *diagnostic*. Nous parlons alors aussi d'*ignorance de décision*.
— Mise à jour des croyances en fonction des observations qui est détectée pendant l'étape d'*extrapolation*. Nous parlons d'*ignorance d'extrapolation*.
— Mise à jour des croyances en fonction de l'action qui est détectée pendant l'étape de *mise à jour*. Nous parlons d' *ignorance de mise à jour*.
— Incohérence avec l'inertie et l'action effectuée par l'agent qui est détectée pendant l'étape de *distorsion*. Nous parlons d'*ignorance de distorsion*.

Nous allons donc créer une caractéristique pour chacune de ces natures.

Ignorance de préférence Si $\varphi_{t'}$ est dans le sous-ensemble M^x_{rev} alors cela veut dire que l'agent a ignoré la croyance $\varphi_{t'}$ car la proposition était incohérente avec une autre croyance ψ au pas de temps t. En d'autres termes, l'agent a préféré ψ à $\varphi_{t'}$ pour retrouver la cohérence. Nous notons alors la caractéristique *préférence* qui retourne vraie si $\varphi_{t'} \in \mathcal{I}^x_t$ a été préférée à une autre croyance par l'agent pour garder la cohérence :

$$\textbf{preference}(\varphi_{t'}, \mathcal{I}^x_t) \Leftrightarrow \ \varphi_{t'} \in M^x_{rev},\ M^x_{rev} \subset \mathcal{I}^x_t$$

Ignorance de décision Si $\varphi_{t'}$ est dans le sous-ensemble M^x_{diag} alors cela veut dire que l'agent a ignoré la croyance $\varphi_{t'}$ pour effectuer une action qui n'était pas cohérente avec ses croyances. L'agent a donc pris une mauvaise décision. Nous notons alors la caractéristique *decision* qui retourne vraie si $\varphi_{t'} \in \mathcal{I}^x_t$ est une ignorance liée à une mauvaise décision :

$$\textbf{decision}(\varphi_{t'}, \mathcal{I}^x_t) \Leftrightarrow \varphi_{t'} \in M^x_{diag},\ M^x_{diag} \subset \mathcal{I}^x_t$$

Ignorance d'extrapolation Si $\varphi_{t'}$ est dans le sous-ensemble M^x_{ext} alors cela veut dire que l'agent a observé une information qui nécessite qu'une proposition soit mise à jour. Ces ignorances sont donc attendues chez un agent rationnel. Nous notons alors la caractéristique *extrapolation* qui retourne vraie si $\varphi_{t'} \in \mathcal{I}^x_t$ est une ignorance d'*extrapolation* :

$$\textbf{extrapolation}(\varphi_{t'}, \mathcal{I}^x_t) \Leftrightarrow \varphi_{t'} \in M^x_{ext},\ M^x_{ext} \subset \mathcal{I}^x_t$$

Ignorance de mise à jour Nous avons vu que si $\varphi_{t'}$ est dans le sous-ensemble M^x_{dist} alors cela veut dire qu'il s'agit peut être d'une *mise à jour* ou d'une *distorsion* du fait que notre algorithme ne peut les différencier (voir section VI.4). Nous proposons donc ici de définir une caractéristique capable de différencier une *mise à jour* par rapport à une *distorsion* . Pour cela, nous allons partir du fait qu'une *mise à jour* ne peut être que l'ignorance d'une proposition d'*inertie* de valeur (comme l'*extrapolation*) et que la proposition dont l'*inertie* change doit être la conséquence des effets de l'action effectuée. Pour déterminer si une proposition $\varphi_{t'}$ ignorée est la conséquence d'une action a_t nous allons utiliser à nouveau

notre opérateur de diagnostic $\mathfrak{M}$ mais de manière différente. Considérons que $keep_t^{(val)}(\varphi) \in M^x_{dist}$, $M^x_{dist} \subset \mathcal{I}^x_t$. Si nous ajoutons $\neg known(\varphi)_{t'}$ à l'état de croyance $B_t^{\mathcal{I}^x}$ nous savons que $B_t^{\mathcal{I}^x}$ est incohérent du fait que $known(\varphi)_{t'}$ est forcément vrai du fait que sa valeur de vérité doit être changé. Ainsi, si nous cherchons une correction en autorisant que l'ignorance des règles de l'effet de l'action a_t alors si nous trouvons une correction possible, cela veut dire que $\varphi_{t'}$ est la conséquence de l'effet de l'action a_t sinon aucune correction ne serait possible. Nous définissons alors la caractéristique *update* qui retourne vraie si $\varphi_{t'}$ est une ignorance de *mise à jour* :

$$\textbf{update}(\varphi_{t'}, \mathcal{I}^x_t) \Leftrightarrow \begin{cases} \varphi_{t'} \in M^x_{dist},\ M^x_{dist} \subset \mathcal{I}^x_t \\ \textbf{isValueInertia}(\varphi_{t'}) \\ keep_t^{(val)}(\psi) = \varphi_{t'} \\ \mathfrak{M}\big(\Phi, screened\big) \neq \emptyset \\ \Phi = B_t^{\mathcal{I}^x} \cup \neg known(\psi_t) \cup effect(a_t) \\ screened = B_t^{\mathcal{I}^x} \cup \neg known(\psi_t) \end{cases}$$

Ignorance de distorsion Nous avons vu que nous pouvons différencier la *mise à jour* de la *distorsion* avec la caractéristique *update*. De ce fait, une ignorance de *distorsion* est simplement une ignorance qui appartient à M^x_{dist} mais qui n'est pas une mise à jour :

$$\textbf{distortion}(\varphi_{t'}, \mathcal{I}^x_t) \Leftrightarrow \begin{cases} \varphi \in M^x_{dist},\ M^x_{dist} \subset \mathcal{I}^x_t \\ \neg\textbf{update}(\varphi_{t'}, \mathcal{I}^x_t) \end{cases}$$

VII.1.3.c Caractéristiques liées aux désirs

L'introduction des désirs dans notre modèle permet de déterminer la valence affective associée à un état de croyance, conformément aux théories de l'évaluation cognitive présentées dans le chapitre II. En effet, les émotions négatives sont généralement associées à des buts non satisfaits. Nous proposons de définir des caractéristiques liées aux désirs de l'agent pour expliquer une erreur de raisonnement par le prisme de l'émotion. Par exemple un agent peut choisir d'ignorer une information qui va à l'encontre de ses buts (c'est la célèbre « pensée désidérative » [Bastardi et al., 2011] présente par exemple dans le biais d'optimisme). Nous allons donc définir des caractéristiques pour déterminer si un état de croyance qui résulte d'un MCS satisfait, ne satisfait pas ou n'infère pas un désir. Nous verrons ensuite que nous pouvons définir des satisfactions fortes et faibles sur les désirs.

De manière plus générale, nous notons les caractéristiques émotionnelles avec deux paramètres $d_{t'}$ qui est un désir (*i.e* $d_{t'} \in \mathcal{D}$) et B^x_t un état de croyances.

Satisfaction d'un désir Un désir $d_{t'}$ est satisfait si, à partir d'un état de croyances B^x_t, l'agent peut déduire par raisonnement le désir $d_{t'}$. Pour vérifier cette propriété nous allons tout simplement retirer $d_{t'}$ de l'état de croyances et vérifier que $d_{t'}$ est toujours une conséquence logique de B^x_t. C'est-à-dire que $B^x_t \vdash d_{t'}$ (voir

sous-section V.2.1). Nous notons alors la caractéristique *satisfied* qui retourne vrai si le désir $d_{t'}$ est satisfait dans B_t^x :

$$\textbf{satisfied}(d_{t'}, B_t^x) \Leftrightarrow (B_t^x \setminus \{d_{t'}\}) \vdash d_{t'}$$

Non-satisfaction d'un désir Un désir $d_{t'}$ n'est pas satisfait si, à partir d'un état de croyances B_t^x, l'agent peut déduire par raisonnement $\neg d_{t'}$. Nous définissons alors la caractéristique *notSatisfied* qui retourne vrai si le désir $d_{t'}$ n'est pas satisfait dans B_t^x de manière similaire que la caractéristique *satisfied* :

$$\textbf{notSatisfied}(d_{t'}, B_t^x) \Leftrightarrow (B_t^x \setminus \{d_{t'}\}) \vdash \neg d_{t'}$$

Non-inférence d'un désir Un désir $d_{t'}$ n'est pas inféré si, à partir d'un état de croyances B_t^x, l'agent ne peut pas déduire $d_{t'}$ ou $\neg d_{t'}$ par raisonnement. Nous définissons alors la caractéristique *noInfer* qui retourne vrai si le désir $d_{t'}$ n'est pas inféré dans B_t^x :

$$\textbf{noInfer}(d_{t'}, B_t^x) \Leftrightarrow \begin{cases} (B_t^x \setminus \{d_{t'}\}) \nvdash d_{t'} \\ (B_t^x \setminus \{d_{t'}\}) \nvdash \neg d_{t'} \end{cases}$$

VII.1.3.d Synthèse

Nous avons donc trois types de caractéristiques intrinsèques selon qu'on s'intéresse au type de ce qui a été ignoré (observations, règles, désirs...), à la nature de l'ignorance (révision de croyance, extrapolation, inertie...) ou à la satisfaction des désirs dans l'état de croyances. L'ensemble est résumé sur le tableau VII.1 avec les caractéristiques et leurs paramètres, classés par type de caractéristiques.

	Type	Nature	Désir
Caractéristiques	inObs isRule isCondition isConsequence isKnownInertia isValueInertia isDesire	preference decision extrapolation update distortion	satisfied notSatisfied noInfer
Paramètres	φ'_t : ignorance t : pas de temps de l'ignorance	$\varphi_{t'}$: ignorance $\mathcal{I}_t^x$: ensemble des ignorances	$d_{t'}$: désir B_t^x : état de croyances

Table VII.1 – Synthèse des caractéristiques intrinsèques

VII.1.4 Caractéristiques locales

Pour définir les caractéristiques locales des ignorances et états de croyances, nous nous plaçons sur une granularité où nous avons accès aux caractéristiques intrinsèques d'un ensemble d'ignorances $\mathcal{I}_t$ mais aussi des ignorances voisines de celui-ci dans l'arbre de diagnostic $\mathfrak{T}$. Nous pouvons donc utiliser la fonction de choix (voir section VI.6) pour définir de nouvelles caractéristiques. Nous distinguons deux types de caractéristiques possibles liées au choix :

— le choix lié aux désirs (*e.g* un désir satisfait est préféré à un désir non satisfait) ;

— le choix de coût cognitif (*e.g* une révision moins coûteuse est préférée par l'agent) ;

Nous allons définir pour chaque type plusieurs caractérisations possibles.

VII.1.4.a Choix lié aux désirs

L'agent peut faire face à une situation où un de ses désirs n'est pas satisfait. Dans ce cas plusieurs *alternatives* (voir section VI.6) sont possibles à un état de croyances : des états de croyances où le désir est satisfait, insatisfait ou non inféré. Par exemple, nous pouvons voir dans le premier pas de temps de l'accident 447 (voir sous-section V.5.3) que l'agent peut avoir des états de croyances où son désir de ne pas être en décrochage n'est pas satisfait (voir sous-section V.5.3) :

$$\mathcal{I}_1^a \quad = \quad \{R_1^j, R_1^g, \mathcal{D}(\neg \, \mathrm{stall}_1)\}$$

ou des états où son désir de ne pas être en décrochage est satisfait, par exemple :

$$\mathcal{I}_1^c \quad = \quad \{\mathrm{buffet}_1, \mathrm{alarm}_1, \mathcal{D}(\neg \, \mathrm{overspeed}_1)\}$$

Nous pouvons donc comparer ces différentes *alternatives* pour définir des caractéristiques sur l'état de croyances analysé, c'est-à-dire l'état résultat des ignorances $\mathcal{I}_x$ considérées. Nous distinguons en effet deux situations possibles :

(1) L'agent a fait le choix d'éviter qu'un de ces désirs soit satisfait ou insatisfait. Cela veut dire que l'agent se trouve dans un état de croyance où un de ces désirs est insatisfait (respectivement satisfait) ou non inféré alors qu'il existait un autre état de croyances alternatif où le même désir était satisfait (respectivement insatisfait). Nous parlons alors d'*aversion* à la perte ou d'*aversion* au gain en fonction de si l'agent a évité un désir insatisfaisable ou satisfaible.

(2) L'agent a fait le choix de croire qu'un de ces désirs soit satisfait ou insatisfait. Cela veut dire que l'agent se trouve dans un état de croyances où un de ces désirs est satisfait (respectivement insatisfait) alors qu'il existait un autre état de croyances alternatif où le même désir était insatisfait (respectivement satisfait) ou non inféré. Nous parlons alors d'*attrait* au gain

ou à la perte en fonction de si l'agent préfère croire à un désir satisfait ou insatisfait.

Contrairement aux caractéristiques intrinsèques sur les désirs définies dans la section précédentes, il ne s'agit pas ici de savoir si un but est satisfait ou non, mais s'il y avait une alternative à cet état de fait que l'agent a choisi d'ignorer. Nous allons pour chaque situation définir les caractéristiques possibles. Pour cela, nous allons les définir par trois paramètres : $d_{t'}$ un désir de l'agent, B_t^x l'état de croyances analysé et B_t^{alt} un état de croyance alternatif résultant d'un autre choix d'ignorances que les ignorances de B_t^x (voir la fonction *alt* section VI.6).

L'aversion Ce cas correspond à l'un des deux cas suivants :

— $\mathbf{noInfer}(d_{t'}, B_t^x)$ ou $\mathbf{satisfied}(d_{t'}, B_t^x)$ et $\mathbf{notSatisfied}(d_{t'}, B_t^{alt})$: le désir n'est pas inféré ou est satisfait dans B_t^x et insatisfait dans l'état alternatif B_t^{alt}.

— $\mathbf{noInfer}(d_{t'}, B_t^x)$ ou $\mathbf{notSatisfied}(d_{t'}, B_t^x)$ et $\mathbf{satisfied}(d_{t'}, B_t^{alt})$: le désir n'est pas inféré ou insatisfait dans B_t^x et satisfait dans l'état alternatif B_t^{alt}.

Nous notons alors les caractéristiques $aversion^{(+)}$ et $aversion^{(-)}$ pour respectivement l'aversion au gain et l'aversion à la perte :

$$\mathbf{aversion}^{(+)}(d_{t'}, B_t^x, B_t^{alt}) \Leftrightarrow \begin{cases} \mathbf{noInfer}(d_{t'}, B_t^x) \text{ ou } \mathbf{notSatisfied}(d_{t'}, B_t^x) \\ \mathbf{satisfied}(d_{t'}, B_t^{alt}) \end{cases}$$

$$\mathbf{aversion}^{(-)}(d_{t'}, B_t^x, B_t^{alt}) \Leftrightarrow \begin{cases} \mathbf{noInfer}(d_{t'}, B_t^x) \text{ ou } \mathbf{satisfied}(d_{t'}, B_t^x) \\ \mathbf{notSatisfied}(d_{t'}, B_t^{alt}) \end{cases}$$

L'attrait Ce cas correspond à l'un des deux cas suivants :

— $\mathbf{satisfied}(d_{t'}, B_t^x)$ et $\mathbf{noInfer}(d_{t'} B_t^{alt})$ ou $\mathbf{notSatisfied}(d_{t'}, B_t^{alt})$: le désir est satisfait dans B_t^x mais non inféré ou non satisfait dans l'état alternatif B_t^{alt}.

— $\mathbf{notSatisfied}(d_{t'}, B_t^x)$ et $\mathbf{noInfer}(d_{t'}, B_t^{alt})$ ou $\mathbf{satisfied}(d_{t'}, B_t^{alt})$: le désir est insatisfait dans B_t^x et non inféré ou satisfait dans l'état alternatif B_t^{alt}.

Nous notons alors les caractéristiques $appeal^{(+)}$ et $appeal^{(-)}$ pour respectivement l'attrait au gain et l'attrait à la perte :

$$\mathbf{appeal}^{(+)}(d_{t'}, B_t^x, B_t^{alt}) \Leftrightarrow \begin{cases} \mathbf{satisfied}(d_{t'}, B_t^x) \\ \mathbf{notSatisfied}(d_{t'}, B_t^{alt}) \text{ ou } \mathbf{noInfer}(d_{t'}, B_t^{alt}) \end{cases}$$

$$\mathbf{appeal}^{(-)}(d_{t'}, B_t^x, B_t^{alt}) \Leftrightarrow \begin{cases} \mathbf{notSatisfied}(d_{t'}, B_t^x) \\ \mathbf{satisfied}(d_{t'}, B_t^{alt}) \text{ ou } \mathbf{noInfer}(d_{t'}, B_t^{alt}) \end{cases}$$

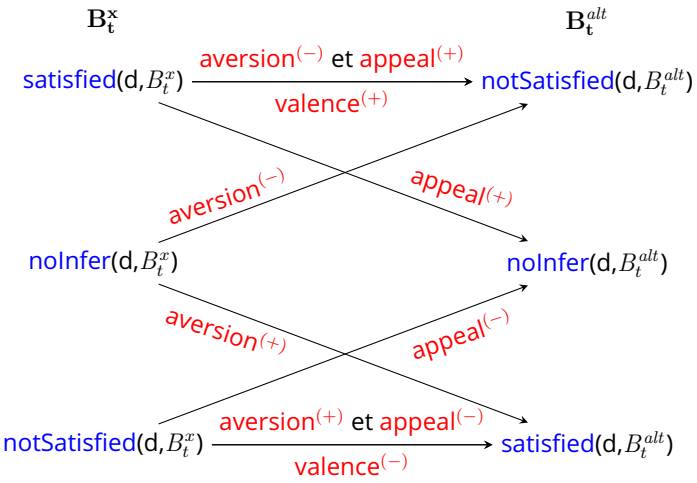


Figure VII.2 – Relation entre l'aversion et l'attrait

Relation entre l'aversion et l'attrait À partir de la définition de l'*aversion* et l'*attrait*, nous pouvons définir une caractéristique qui lie l'*attrait* et l'*aversion*. En effet, dans le cas où l'état de croyances de l'agent peut être caractérisé par un attrait vers un désir et une aversion à la perte de ce même désir : $\mathbf{aversion}^{(-)}(d_{t'}, B_t^x, B_t^{alt})$ et $\mathbf{appeal}^{(+)}(d_{t'}, B_t^x, B_t^{alt})$ alors cela veut dire que l'agent a fait le choix d'un désir satisfait au lieu du même désir insatisfait car :

$$(\mathbf{aversion}^{(-)}(d_{t'}, B_t^x, B_t^{alt}) \text{ et } \mathbf{appeal}^{(+)}(d_{t'}, B_t^x, B_t^{alt})$$
$$\text{est équivalent à}$$
$$(\mathbf{satisfied}(d_{t'}, B_t^x) \text{ et } \mathbf{notSatisfied}(d_{t'}, B_t^{alt}))$$

En d'autres termes, en faisant ce choix de satisfaire son désir, l'agent a évité que son désir ne soit pas satisfait et a eu un attrait pour que son désir soit satisfait. Nous caractérisons ce choix comme une *valence* au gain et le définissons par :

$$\mathbf{valence}^{(+)}(d_{t'}, B_t^x, B_t^{alt}) \Leftrightarrow \begin{cases} \mathbf{aversion}^{(-)}(d_{t'}, B_t^x, B_t^{alt}) \\ \mathbf{appeal}^{(+)}(d_{t'}, B_t^x, B_t^{alt}) \end{cases}$$

De manière analogue, nous définissons la *valence* à la perte par le fait que l'agent a évité que son désir soit satisfait et a eu un attrait pour que son désir soit insatisfait :

$$\mathbf{valence}^{(+)}(d_{t'}, B_t^x, B_t^{alt}) \Leftrightarrow \begin{cases} \mathbf{aversion}^{(+)}(d_{t'}, B_t^x, B_t^{alt}) \\ \mathbf{appeal}^{(-)}(d_{t'}, B_t^x, B_t^{alt}) \end{cases}$$

Nous résumons ces relations par le schéma de relation figure VII.2.

VII.1.4.b Choix du coût

Notre fonction de choix permet de déterminer quelles ignorances pouvaient être choisies à la place d'une ignorance φ'_t dans un ensemble d'ignorances $\mathcal{I}_t^x$

(voir section VI.6) pour retrouver la cohérence. Nous pouvons ensuite à partir de ce choix calculer l'ensemble alternatif d'ignorances $\mathcal{I}_t^{alt}$ où ce choix a été fait et est le plus proche de $\mathcal{I}_t$ grâce à la fonction alt. Nous pouvons donc comparer la taille de ces ensembles d'ignorances. Nous faisons l'hypothèse que plus un agent ignore des propositions logiques, plus le coût cognitif est élevé. Par conséquent, nous pouvons comparer le coût cognitif de chaque étape de notre algorithme. Par exemple, si $|M_{rev}^x| < |M_{rev}^{alt}|$, c'est-à-dire que l'étape de révision à moins d'ignorance que l'étape d'ignorance de l'ensemble alternatif alors, nous pouvons en conclure que la révision de croyance que l'agent a décidé d'effectuer dans $\mathcal{I}_t^x$ est moins couteuse que la révision alternative dans $\mathcal{I}_t^{alt}$.

Par exemple, si nous reprenons une variation de l'exemple de l'alarme et des nuages précédents avec le fait que l'agent peut observer cette fois du tonnerre en plus :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\mathrm{alarm}_1, \neg\,\mathrm{cloud}_1, \mathrm{thunder}_1\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \mathrm{alarm}_t \rightarrow \mathrm{storm}_{t+1} \\ R^b & \equiv & \mathrm{cloud}_t \rightarrow \neg\,\mathrm{storm}_{t+1} \\ R^c & \equiv & \mathrm{thunder}_t \rightarrow \mathrm{storm}_{t+1} \end{array} \right. \end{array}$$

Nous avons comme ignorances possibles :

$$\begin{array}{lcl} \mathcal{I}_1^a & \equiv & \{\mathrm{alarm}_1, R_1^c\} \\ \mathcal{I}_1^b & \equiv & \{\mathrm{alarm}_1, \mathrm{thunder}_1\} \\ \mathcal{I}_1^c & \equiv & \{\mathrm{thunder}_1, R_1^a\} \\ \mathcal{I}_1^d & \equiv & \{R_1^a, R_1^c\} \\ \mathcal{I}_1^e & \equiv & \{\neg\,\mathrm{cloud}_1\} \\ \mathcal{I}_1^f & \equiv & \{R_1^b\} \end{array}$$

Si nous regardons les choix de préférence de alarm_1 pour $\mathcal{I}_1^a$, c'est-à-dire $choice(\mathrm{alarm}_1, M_{rev}^a)$ nous trouvons :

$$\{R_1^a\}, \{\neg\,\mathrm{cloud}_1\}, \{R_1^b\}$$

De ces choix, nous trouvons comme alternatives pour $\mathcal{I}_1^a$:

$$\begin{array}{lcl} alt(\{R_1^a\}, \mathcal{I}_1^a) & = & \mathcal{I}_1^d \\ alt(\{\neg\,\mathrm{cloud}_1\}, \mathcal{I}_1^a) & = & \mathcal{I}_1^e \\ alt(\{R_1^b\}, \mathcal{I}_1^a) & = & \mathcal{I}_1^f \end{array}$$

Nous pouvons alors en déduire que le choix d'ignorer R_1^a à la place de alarm_1 dans $\mathcal{I}_1^a$ est un choix de révision au coût équivalent car $M_d^{rev} = \{R_1^a, R_1^c\}$ et $M_{rev}^a = \{\mathrm{alarm}_1, R_1^c\}$. Toutefois, le choix d'ignorer R_1^b ou $\neg\,\mathrm{cloud}_1$ est moins couteux du fait que $M_{rev}^e = \{\neg\,\mathrm{cloud}_1\}$ et $M_{rev}^f = \{R_1^b\}$.

Nous définissons trois caractéristiques permettant de déterminer si un choix est plus ou moins coûteux pour une étape. Nous allons définir ces caractéristiques par trois paramètres : $\mathcal{I}_t^x$ l'ensemble des ignorances analysé et $\mathcal{I}_t^{alt}$ une alternative à l'ensemble $\mathcal{I}_t^x$ et $step$ l'étape de l'algorithme de diagnostic.

Choix au coût équivalent Pour un choix au coût équivalent, nous notons la caractéristique $equalCostly$:

$$\textbf{equalCostly}(\mathcal{I}_t^x, \mathcal{I}_t^{alt}, step) \Leftrightarrow \ |M_{step}^x| = |M_{step}^{alt}|$$

Choix coûteux Pour un choix plus coûteux, nous notons la caractéristique $costly$:

$$\textbf{costly}(\mathcal{I}_t^x, \mathcal{I}_t^{alt}, step) \Leftrightarrow \ |M_{step}^x| > |M_{step}^{alt}|$$

Choix moins coûteux Pour un choix moins coûteux, nous notons la caractéristique $lessCostly$:

$$\textbf{lessCostly}(\mathcal{I}_t^x, \mathcal{I}_t^{alt}, step) \Leftrightarrow \ |M_{step}^x| < |M_{step}^{alt}|$$

VII.1.4.c Synthèse

Nous avons donc deux types de caractéristiques locales selon qu'on s'intéresse au choix de l'agent par rapport au désir (satisfait, insatisfait ou non inféré) ou au coût d'un autre choix d'ignorance en fonction du nombre d'ignorances faites dans ce choix alternatif. L'ensemble est résumé sur le tableau VII.2 avec les caractéristiques et leurs paramètres.

	Désir	Coût
Caractéristiques	aversion$^{(+)}$ aversion$^{(-)}$ appeal$^{(+)}$ appeal$^{(-)}$ valence$^{(+)}$ valence$^{(-)}$	equalCostly costly lessCostly
Paramètres	B_t^x : état de croyances B_t^{alt} : état alternatif	$\mathcal{I}_t^x$: ensemble d'ignorances $\mathcal{I}_t^{alt}$: ensemble alternatif $step$: étape de l'algorithme

Table VII.2 – Synthèse des caractéristiques locales

VII.1.5 Caractéristiques globales

Pour les caractéristiques globales, nous nous plaçons sur une granularité où nous avons accès, en plus des caractéristiques intrinsèques et locales, à tout l'historique des ignorances d'un scénario dans lequel un ensemble d'ignorance $\mathcal{I}_t^x$ appartient. Ces caractéristiques permettent d'analyser des propriétés à travers le temps ainsi que de définir dans quel contexte se trouve l'état de croyances analysé. Par exemple, est-ce que l'agent a choisi d'ignorer l'information la plus ancienne ? Ou encore est-ce que l'agent est passé d'un état où son désir n'était pas satisfait à un état où

il est satisfait ? Nous pouvons distinguer trois types de caractéristiques sur cette granularité :

— les choix temporels qui correspondent à des choix en fonction de l'ancienneté des croyances ;
— les similitudes entre les états de croyances passés du scénario qui sont des motifs répétés dans le scénario ;
— les changements des désirs entre les états de croyances passés du scénario ;

Nous allons donc définir dans les prochaines sections les différentes caractéristiques de ces trois types. Afin de faciliter les notations des caractéristiques *globales* nous ne noterons pas les indices de chemin de corrections sur les états de croyances. En effet, les caractéristiques globales permettent de comparer un état B_t^x aux autres états parents de B_t^x. Nous noterons donc B_t comme un état de croyance quelconque et B_i avec $i \neq t$ comme un état de croyance parent de B_t.

Toutefois, cette analyse à travers le temps des croyances de l'agent nous pousse à définir quels sont les croyances qui sont équivalentes dans le temps afin de définir les différentes caractéristiques. Nous allons dans un premier temps définir un opérateur qui permet de déterminer si une croyance était crû dans un état de croyances précédent.

VII.1.5.a Opérateur de croyance précédente

Déterminer si une croyance $\varphi_{t'}$ était crue à un instant i précédemment à un instant t n'est pas aussi simple que de déterminer si $B_i \vdash \varphi_{t'}$. En effet, en fonction de l'indice t' de φ nous devons considérer plusieurs cas. Considérons l'exemple suivant avec les observations :

$$\begin{aligned} Obs_1 &\equiv \{\mathrm{alarm}_1, \mathrm{cloud}_1\} \\ Obs_2 &\equiv \{\mathrm{alarm}_2, \mathrm{cloud}_3\} \end{aligned}$$

Ici l'agent peut observer des nuages et une alarme au temps 1, puis, au temps 2, il peut observer à nouveau une alarme et une prévision de nuages pour le temps 3. Il y a donc des croyances de différents types : des prévisions quand l'indice temporel de la proposition est supérieur au pas de temps courant et des croyances sur l'état actuel du monde quand la proposition a un indice temporel équivalent au pas de temps courant. Nous considérons que les croyances sur l'état actuel sont les mêmes informations que les croyances sur l'état du monde passé du fait de l'*inertie*. Dans l'exemple, la croyance sur l'alarme au pas de temps 1 est la même que celle au pas de temps 2 par l'*inertie* (*i.e* $\mathrm{alarm}_t \longleftrightarrow \mathrm{alarm}_{t-1}$) : l'agent observe finalement le même état du monde sur deux pas de temps différents. Ce n'est pas le cas pour les prévisions car par *inertie* l'observation de cloud_1 mène à croire au pas de temps 2 cloud_2 qui est donc différent de la croyance cloud_3.

Nous notons alors l'opérateur *previousBelief* qui retourne vrai si une croyance φ'_t au temps t est aussi une croyance (la même information) dans l'état de croyance

précédent B_i :

$$\textbf{previousBelief}(\varphi_{t'}, t, B_i) \Leftrightarrow \begin{cases} i < t \\ B_i \vdash \varphi_{t'} \text{ ou } (t = t' \text{ et } B_i \vdash \varphi_i)) \end{cases}$$

VII.1.5.b Choix temporel

L'agent peut choisir d'ignorer une proposition plutôt qu'une autre en fonction de son ancienneté. Prenons par exemple un agent qui peut observer au pas de temps 1 que la météo lui indique qu'il n'y aura pas d'orage au pas de temps 3 puis observe plus tard du tonnerre au loin qui lui indique un orage au pas de temps 3. Nous avons alors :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\neg \text{storm}_3\} \\ Obs_2 & \equiv & \{\text{thunder}_2\} \\ \mathcal{R} & \equiv & \{ \ R^a \ \equiv \ \text{thunder}_t \rightarrow \text{storm}_{t+1} \end{array}$$

Dans ce cas, l'agent peut choisir par exemple soit d'ignorer la prévision météo du temps 1 ou le tonnerre du temps 2. Il y a donc deux situations possibles :

— l'agent préfère garder ses anciennes croyances ;

— l'agent préfère garder les nouvelles croyances.

Nous considérons que si une croyance $\varphi_{t'}$ est crue dans l'état de croyance précédent, c'est une croyance ancienne, sinon c'est une nouvelle croyance. Pour capturer le fait qu'un agent préfère ses anciennes croyances, nous définissons la caractéristique de *conservatisme* qui prend en paramètre une ignorance $\varphi_{t'}$, un autre choix d'ignorance ψ_i pour $\varphi_{t'}$ et l'état de croyance précédent B_{t-1} :

$$\textbf{conservatisme}(\varphi_{t'}, \psi_i, B_{t-1}) \Leftrightarrow \begin{cases} \neg\textbf{previousBelief}(\varphi_{t'}, t, B_{t-1}) \\ \textbf{previousBelief}(\psi_i, t, B_{t-1}) \end{cases}$$

En d'autres termes, si l'agent ignore une nouvelle croyance $\varphi_{t'}$ (*i.e* n'est pas une croyance dans l'étape précédente) à la place d'une croyance ψ_i de l'état précédent alors l'agent a préféré ses anciennes croyances.

Enfin, de manière analogue nous notons $credulity$ qui retourne vraie si l'agent préfère une nouvelle croyance par rapport à une ancienne croyance :

$$\textbf{credulity}(\varphi_{t'}, \psi_i, B_{t-1}) \Leftrightarrow \begin{cases} \textbf{previousBelief}(\varphi_{t'}, t, B_{t-1}) \\ \neg\textbf{previousBelief}(\psi_i, t, B_{t-1}) \end{cases}$$

VII.1.5.c Similitudes entre états de croyances

Un état de croyances peut avoir un contexte similaire à un état de croyances dans le passé, c'est-à-dire, une action, des observations, des désirs et/ou des croyances similaires. La littérature montre que de nombreux biais proviennent de cette similitude de situation. En effet, lorsque l'agent fait face à des situations

proches, la confiance en son jugement peut être plus élevé du fait qu'il a déjà rencontré la situation (*e.g* l'excès de confiance, l'illusion de contrôle, etc). Nous allons donc définir des caractéristiques pour la similitude pour les observations, actions et désirs.

Similitudes des observations Une même information peut être observée de nombreuses fois dans les états mentaux passés et ainsi renforcer du point de vue de l'agent la vérité de cette information par rapport à un autre choix de proposition. Pour déterminer si une information $\varphi_{t'}$ peut être observée dans un état de croyances précédent B_i, nous allons vérifier que $\varphi_{t'}$ est cru dans B_i et B_t et fait partie des observations du pas de temps i et t. Nous notons la caractéristique *sameObs* qui retourne vrai si l'information $\varphi_{t'}$ est observé dans B_t et B_i :

$$\textbf{sameObs}(\varphi_{t'}, B_t, B_i) \Leftrightarrow \begin{cases} i < t \\ \textbf{inObs}(\varphi_{t'}, t) \\ B_t \vdash \varphi_{t'} \\ \textbf{inObs}(\varphi_{t'}, i) \text{ ou } (t' = t \text{ et } \textbf{inObs}(\varphi_i, i)) \\ \textbf{previousBelief}(\varphi_{t'}, t, B_i) \end{cases}$$

Similarité des actions Une action répétée plusieurs fois peut être un indice intéressant pour expliquer les croyances de l'agent. Par exemple, si l'agent a répété la même action et la même ignorance au même pas de temps t, la raison de l'ignorance à de grande chance d'être similaire que la première fois. Nous notons la qualité *sameAction* qui retourne vrai si l'action au pas de temps t a été déjà fait à l'instant précédent i :

$$\textbf{sameAction}(B_t, B_i) \Leftrightarrow \begin{cases} i < t \\ \exists a'_i, a_t \in \mathcal{A}, a'_i = a_i \end{cases}$$

Similarité des désirs Un même type de désir qui se retrouve insatisfait ou satisfait à un pas de temps t différent permet potentiellement d'expliquer que l'agent se retrouve dans un contexte similaire que précédemment et par conséquent expliquer sa décision pour satisfaire ou non ce désir. Nous notons alors la caractéristique $sameDesire^{(+)}$ qui retourne vraie si le même désir $d_{t'}$ est satisfait dans B_t et B_i :

$$\textbf{sameDesire}^{(+)}(\varphi_{t'}, B_t, B_i) \Leftrightarrow \begin{cases} i < t \\ \textbf{isDesire}(\varphi_{t'}) \\ \textbf{satisfied}(\varphi_{t'}, B_t) \\ \textbf{satisfied}(\varphi_{t'}, B_i) \text{ ou } (t' = t \text{ et } \\ \qquad\qquad \textbf{satisfied}(\varphi_i, B_i)) \end{cases}$$

Nous notons la caractéristique $sameDesire^{(-)}$ qui retourne vraie si le même désir $d_{t'}$ n'est pas satisfait dans B_t et B_i :

$$\textbf{sameDesire}^{(-)}(\varphi_{t'}, B_t, B_i) \Leftrightarrow \begin{cases} i < t \\ \textbf{isDesire}(\varphi_{t'}) \\ \textbf{notSatisfied}(\varphi_{t'}, B_t) \\ \textbf{notSatisfied}(\varphi_{t'}, B_i) \text{ ou } (t' = t \text{ et } \\ \qquad\qquad \textbf{notSatisfied}(\varphi_i, B_i)) \end{cases}$$

VII.1.5.d Changements des désirs entre les états de croyances

D'un état de croyances B_{t-1} à un état B_t il est possible que certains désirs de l'agent changent de valeur de vérité. Ce changement peut être une partie de l'explication du choix d'une ignorance et/ou d'une action. En effet, si l'agent croit qu'une action a eu pour effet de changer l'état (satisfait ou non) d'un désir dans le passé, il peut décider d'effectuer des décisions similaires ou au contraire éviter de refaire les mêmes erreurs. Par exemple, si l'agent prend la décision de tirer le manche lors d'un décrochage et observe que l'alarme de décrochage ne sonne plus, alors il est probable qu'il prenne la même décision si l'alarme sonne une nouvelle fois s'il désire ne pas être en décrochage. Nous distinguons donc deux types de caractéristiques de changements à prendre en compte :

— Un changement positif pour l'agent : un désir passe de non satisfait à satisfait.
— Un changement négatif pour l'agent : un désir passe de satisfait à non satisfait.

Changement positif Deux situations sont à prendre en compte pour un changement positif pour l'agent :

— L'agent croit au temps $t-1$ que $d_{t'}$ n'est pas satisfait mais croit au temps t que $d_{t'}$ est satisfait. C'est un changement dû à une *révision*, car l'agent était face à une contradiction et a préféré croire que son désir était satisfait.
— L'agent croit au temps $t-1$ que d_{t-1} n'est pas satisfait mais croit au temps d_t est satisfait. C'est un changement dû à une *mise à jour*, car le monde a changé et son désir est maintenant satisfait.

Nous notons alors la caractéristique $changeRevision^{(+)}$ qui retourne vrai si le désir $d_{t'}$ entre le pas de temps t et $t-1$ a subit un changement positif dû à une *révision* :

$$\mathbf{changeRevision}^{(+)}(d_{t'}, B_t, B_{t-1}) \Leftrightarrow \begin{cases} \mathbf{isDesire}(d_{t'}) \\ \mathbf{satisfied}(d_{t'}, B_t) \\ \mathbf{notSatisfied}(d_{t'}, B_{t-1}) \end{cases}$$

et pour un changement positif dû à une *mise à jour*, nous notons :

$$\mathbf{changeUpdate}^{(+)}(d_t, B_t, B_{t-1}) \Leftrightarrow \begin{cases} \mathbf{isDesire}(d_{t'}) \\ \mathbf{satisfied}(d_t, B_t) \\ \mathbf{notSatisfied}(d_{t-1}, B_{t-1}) \end{cases}$$

Changement négatif De manière analogue au changement positif, nous capturons le changement négatif d'un désir $d_{t'}$ entre le pas de temps t et $t-1$ dû à une *révision* par :

$$\mathbf{changeRevision}^{(-)}(d_{t'}, B_t, B_{t-1}) \Leftrightarrow \begin{cases} \mathbf{isDesire}(d_{t'}) \\ \mathbf{notSatisfied}(d_{t'}, B_t) \\ \mathbf{satisfied}(d_{t'}, B_{t-1}) \end{cases}$$

et pour un changement négatif dû à une *mise à jour*, nous notons :

$$\mathbf{changeUpdate}^{(-)}(d_{t'}, B_t, B_{t-1}) \Leftrightarrow \begin{cases} \mathbf{isDesire}(d_{t'}) \\ \mathbf{notSatisfied}(d_t, B_t) \\ \mathbf{satisfied}(d_{t-1}, B_{t-1}) \end{cases}$$

VII.1.5.e Synthèse

Nous avons donc trois types de caractéristiques globales selon qu'on s'intéresse au choix temporel de l'agent (préférer ou non les anciennes croyances), les similitudes entre les états de croyances passées et l'état de croyances présent et les changements de satisfaction d'un désir entre deux états de croyances successifs. L'ensemble de ces caractéristiques est résumé sur le tableau VII.3 avec leurs paramètres correspondants.

	Choix temporel	Similitudes	Changements
Caractéristiques	conservatisme credulity	sameObs sameAction sameDesire$^{(+)}$ sameDesire$^{(-)}$	changeRevision$^{(+)}$ changeRevision$^{(-)}$ changeUpdate$^{(+)}$ changeUpdate$^{(-)}$
Paramètres	B_t : état de croyances B_t^{alt} : état alternatif	$\mathcal{I}_t^x$: ensemble d'ignorances $\mathcal{I}_t^{alt}$: ensemble alternatif	$d_{t'}$: un désir B_t : état de croyances B_{t-1} : état de croyances précédent

Table VII.3 – Synthèse des caractéristiques globales

VII.1.6 Conclusion

Nous avons pu définir dans cette section un ensemble de caractéristiques formelles sous forme de fonctions booléennes. Ces caractéristiques sont des propriétés sur les ignorances et états de croyances d'un scénario qui permettent de les différencier. Ces caractéristiques sont définies sur trois granularités d'analyses :

- *intrinsèque* qui sont les caractéristiques propres aux ignorances et état de croyances d'un pas de temps t d'un scénario donné ;
- *locale* qui sont les caractéristiques en comparaison aux choix d'ignorances et alternatives d'états de croyances au pas de temps t ;
- *globale* qui sont les caractéristiques en comparaison aux états de croyances précédents du scénario donné.

Le choix pour une ignorance φ correspond à un ensemble de propositions Ψ qui aurait pu être ignoré à la place de φ et une alternative correspond à un état de croyances où Ψ a été corrigé et φ gardé.

L'ensemble de ces caractéristiques nous permettent de définir des propriétés sur lesquels nous allons nous reposer pour définir les biais cognitifs.

VII.2 Définition des biais

Nous avons vu sous-section IV.2.5 que notre approche consiste à définir un biais cognitif comme une fonction booléenne prenant en paramètre un état de croyances et une ignorance et retourne vraie si un ensemble de caractéristique sont respectées. Nous allons dans cette section définir plusieurs biais cognitifs selon ce principe avec un exemple d'illustration à chaque fois. Cette section n'a pas pour objectif d'être exhaustive : nous avons choisis quelques biais représentatifs des situations d'accident que nous avons étudié au cours de la thèse. Toutefois, nous présentons une méthodologie de définition des biais qui pourrait être étendue à un grand nombre de biais.

VII.2.1 Biais de confirmation

Le biais de confirmation est la tendance à privilégier les informations qui confirment nos croyances et à donner moins d'importance aux informations contradictoires [Nickerson, 1998]. D'un point de vue plus formel, nous considérons qu'il existe un biais de confirmation au pas de temps t :

(1) l'agent fait face à deux informations contradictoires ψ_i et $\varphi_{t'}$;
(2) l'agent préfère l'information ψ_i qui confirme ses anciennes croyances B_{t-1} ;
(3) l'agent ignore $\varphi_{t'}$ ou tout raisonnement utilisant $\varphi_{t'}$.

De cette définition, nous en concluons que nous sommes face à un problème de préférence du fait que l'agent peut observer deux informations contradictoires (ψ_i et $\varphi_{t'}$) et préfère en garder une plutôt qu'une autre (garder ψ_i). Nous recherchons donc une caractéristique d'*ignorance de cohérence* que nous avons définie sous-section VII.1.3. De plus, nous devons caractériser l'autre choix d'ignorance pour déterminer si ce choix "confirme les anciennes croyances". Nous devons donc définir formellement ce que nous entendons par confirmer les anciennes croyances.

Nous définissons le fait qu'une information ψ_i confirme les anciennes croyances par le fait que ψ_i et certaines anciennes croyances dans B_{t-1} doivent aussi permettre de déduire la même chose. Par conséquent, si $\psi_i \rightarrow \neg\varphi_{t'}$ (*i.e* ψ_i est en contradiction avec $\varphi_{t'}$) alors les anciennes croyances doivent aussi permettre de déduire $\neg q_i$ aussi (*i.e* sont en contradiction avec $\varphi_{t'}$ aussi). Toutefois, les anciennes croyances peuvent déduire $\neg\varphi_{t'}$ de deux façons :

(a) Soit $B_{t-1} \vdash \neg\varphi_{t'}$, c'est-à-dire que l'agent pouvait déduire directement de ses croyances précédentes $\neg\varphi_{t'}$. Par conséquent, si l'agent choisit de garder $\varphi_{t'}$ dans ses croyances, il doit ignorer ψ_i ainsi que des croyances précédentes à la place de $\varphi_{t'}$. En d'autres termes, c'est une révision de croyance plus couteuse alors que l'agent avait préféré garder ses anciennes croyances en gardant ψ_i : choisir d'ignorer $\varphi_{t'}$ est donc du conservatisme (voir sous-section VII.1.5). Nous devons donc vérifier qu'il existe un choix pour ψ_i tel qu'il existe dans ce choix une ignorance d'une observation du pas de temps

t et une autre croyance ignorée qui correspond à du conservatisme par rapport à ψ_i. Nous notons l'opérateur $conserv^{(C)}$ qui prend en paramètre une ignorance $\varphi_{t'}$, un ensemble $\mathcal{I}_t^x$ d'ignorance tel que $\varphi_{t'} \in \mathcal{I}_t^x$ retourne vrai si l'ignorance $\varphi_{t'}$ permet de garder une observation qui confirme les anciennes croyances :

$$\mathbf{confirm}^{(\mathbf{C})}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \exists \Psi \in choice(\varphi_{t'}, M_{rev}^x) \\ \exists \psi_i, \gamma_j \in \Psi \\ \exists \mathcal{I}_{t-1}^y \in S(\mathcal{I}_t^x) \\ \mathbf{inObs}(\psi_i, t) \\ \mathbf{conservatisme}(\varphi_{t'}, \gamma_j, B_{t-1}^y) \end{cases}$$

(b) Soit $\varphi_{t'} = \varphi_t$ et $B_{t-1} \vdash \neg\varphi_{t-1}$. Par conséquent, par inertie l'agent peut déduire $\neg\varphi_t$. Cela veut dire que si l'agent choisit de garder $\varphi_{t'}$ dans ses croyances, il doit ignorer ψ_i à la place de φ_t et mettre à jour par extrapolation les croyances précédentes qui sont en contradiction avec φ_t. Si ψ_i confirme ses anciennes croyances, la mise à jour d'extrapolation en gardant φ_t ne peut être que plus couteuse que la mise à jour d'extrapolation en gardant ψ_i. En effet, si ψ_i permet de déduire la même chose que les anciennes croyances, il y a moins de chose à mettre à jour que φ_t qui est en contradiction avec les anciennes croyances. Nous devons donc vérifier que si la proposition ignorée à un indice temporel équivalent à t , il existe un autre choix d'ignorances qui ignore une observation et a pour conséquence une mise à jour par extrapolation plus couteuse.

$$\mathbf{confirm}^{(\mathbf{I})}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} t' = t \\ \exists \Psi \in choice(\varphi_{t'}, M_{rev}^x) \\ \exists \psi_i \in \Psi \\ \mathbf{inObs}(\psi_i, t) \\ \mathbf{costly}(\mathcal{I}_t^x, \mathcal{I}_t^{alt}, ext) \end{cases}$$

avec $\mathcal{I}_t^{alt} = alt(\Psi, \mathcal{I}_x^t)$

.

Nous notons donc la fonction *confirmation* qui retourne vraie si l'ignorance $\varphi_{t'}$ dans les ignorances $\mathcal{I}_t^x$ représente un biais de confirmation :

$$\mathbf{confirmation}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \mathbf{preference}(\varphi_{t'}, \mathcal{I}_t^x) \\ \mathbf{inObs}(\varphi_{t'}, t) \text{ ou } \mathbf{isRule}(\varphi_{t'}) \\ \mathbf{confirm}^{(\mathbf{C})}(\varphi_{t'}, \mathcal{I}_t^x) \text{ ou } \mathbf{confirm}^{(\mathbf{I})}(\varphi_{t'}, \mathcal{I}_t^x) \end{cases}$$

Exemple Pour illustrer ce biais, prenons un exemple que nous allons limiter à un problème de cohérence entre deux informations afin de faciliter la compréhension :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\text{greenLight}_1\} \\ Obs_2 & \equiv & \{\text{reserve}_2, \text{alarm}_2\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \text{greenLight}_t \Rightarrow \neg\,\text{outFuel}_t \\ R^b & \equiv & \text{reserve}_t \Rightarrow \neg\,\text{outFuel}_t \\ R^c & \equiv & \text{alarm}_t \Rightarrow \text{outFuel}_t \end{array} \right\} \end{array}$$

Ici l'agent peut voir au temps 1 un voyant vert qui lui indique qu'il a assez d'essence pour continuer. Toutefois, au pas de temps 2 l'agent peut observer à la fois une alarme qui lui indique qu'il n'a plus d'essence pour continuer et sa réserve qui est plein d'essence. Si nous recherchons les ignorances au pas de temps 2, nous trouvons :

$$\begin{array}{lcl} \mathcal{I}_2^a & = & \{R_2^c\} \\ \mathcal{I}_2^b & = & \{\mathrm{alarme}_2\} \\ \mathcal{I}_2^c & = & \{\mathrm{reserve}_2, keep_{greenLight}^{(val)}(2)\} \\ \mathcal{I}_2^d & = & \{R_2^b, keep_{greenLight}^{(val)}(2)\} \end{array}$$

Nous cherchons à savoir si le fait d'ignorer la proposition alarme_2 dans l'ensemble $\mathcal{I}_2^b$ équivaut à un biais de confirmation. La proposition fait bien partie du problème de préférence du fait qu'aucune action ne rentre en jeu dans l'incohérence. De plus la proposition est une observation du temps 2, la qualité *inObs* est donc respectée. Nous trouvons comme choix possible pour $choice(\mathrm{alarm}_2, M_{rev}^b)$:

$$\{\{R_2^c\}, \{R_2^b\}, \{\mathrm{reserve}_2\}\}$$

Il existe donc une autre observation ($\mathrm{reserve}_2$) qui est incohérente avec l'alarme. De plus, nous trouvons comme alternative avec ce choix : $alt(\{\mathrm{reserve}_2\}, \mathcal{I}_2^b) = \mathcal{I}_2^c$ où la mise à jour par extrapolation est plus coûteuse du fait que $keep_{greenLight}^{(val)}(2)$ est ignoré alors que dans $\mathcal{I}_2^b$ ne contient pas d'ignorance de mise à jour. Ainsi l'opérateur $confir^{(I)}$ est vrai ainsi que la caractéristique $confirmation$ qui à toutes ces conditions respectées. Par conséquent, ignorer l'alarme revient à un biais de confirmation dans cet exemple.

VII.2.2 Biais d'attention

Le biais d'attention est la tendance à sélectionner des informations selon les préoccupations de la personne. Plusieurs types d'attention sont à différencier [Cisler et al., 2010] :

- biais de facilitation : l'attention vers un danger est facilitée. Par exemple, un pilote qui désire ne pas être en panne d'essence va concentrer son attention sur les indices indiquant une panne d'essence en omettant d'autres informations.
- biais de désengagement : le désengagement de l'attention envers un danger est difficile. Par exemple, un pilote qui voit son attention facilitée vers un danger peut difficilement défaire son attention de ce danger tant qu'il est présent. Par conséquent, son attention est facilitée sur plusieurs pas de temps.
- biais d'évitement : l'attention peut être dirigée vers une information opposée au danger. Par exemple, une personne arachnophobe concentrera son attention sur autre chose que l'araignée.

Nous présentons la modélisation de ces trois biais à partir de nos caractéristiques.

VII.2.2.a Biais de facilitation

D'un point de vue logique, l'attention facilitée vers un danger peut être exprimée par deux cas distincts :

— Ce que nous appelons la *facilitation de préférence* qui consiste à avoir un choix entre deux observations incohérentes entre elles et que l'agent choisisse de garder dans ses croyances celle qui mène à l'insatisfaction d'un désir. Par exemple, lorsque les pilotes du vol Rio-Paris se sont concentrés sur les informations de survitesse.

— Ce que nous appelons la *facilitation de décision* qui correspond au fait que l'agent effectue une action incohérente du fait que son attention était concentrée sur une observation qui implique un non-désir. Par exemple, lorsque les pilotes du vol du mont Sainte-Odile se sont concentrés sur l'alignement avec la piste.

Nous allons formaliser ces deux notions.

Facilitation de préférence Pour le cas de la *facilitation de préférence*, nous considérons que nous sommes dans une situation où :

(1) $\varphi_{t'}$ et ψ_i sont deux observations en contradiction
(2) ψ_i est en contradiction avec un désir d_j mais $\varphi_{t'}$ ne l'est pas
(3) $\varphi_{t'}$ est ignoré

De cette définition, nous en concluons que nous sommes face à un problème de préférence du fait que l'agent peut observer deux informations contradictoires (ψ_i et $\varphi_{t'}$) et préfère en garder une plutôt qu'une autre (garder ψ_i). Nous recherchons donc une caractéristique d'*ignorance de préférence* que nous avons définie sous-section VII.1.3. De plus, en préférant ψ_i à $\varphi_{t'}$, l'agent a décidé de se retrouver dans un état de croyance où son désir d_j n'est pas satisfait alors qu'il pouvait ne pas l'être en préférant $\varphi_{t'}$. L'agent a donc un attrait pour une perte, c'est pourquoi nous recherchons une caractéristique d'attrait que nous avons défini sous-section VII.1.4. Nous notons l'opérateur $focus^{(-)}$ qui retourne vraie si l'agent s'est concentré pour l'étape $step$ de l'algorithme de diagnostic sur un autre choix d'observation que $\varphi_{t'}$ qui permet de déduire $\neg d_j$:

$$\mathbf{focus}^{(-)}(\varphi_{t'}, d_j, \mathcal{I}_t^x, step) \Leftrightarrow \begin{cases} \exists \Psi \in choice(\varphi_{t'}, M_{step}^x),\ \exists \psi_i \in \Psi,\ \mathbf{inObs}(\psi_i, t) \\ \psi_i \notin \mathcal{I}_t^x \\ \mathbf{appeal}^{(-)}(d_j, B_t^x, B_t^{alt}) \end{cases}$$

avec $\mathcal{I}_t^{alt} = alt(\Psi, \mathcal{I}_t^x)$

Nous notons alors le biais $facilitation^{(P)}$ qui retourne vrai si l'ignorance de $\varphi_{t'}$ dans l'ensemble $\mathcal{I}_t^x$ correspond à un biais de facilitation de cohérence sur le désir d_j :

$$\mathbf{facilitation}^{(\mathbf{P})}(\varphi_{t'}, \mathcal{I}_t^x, d_j) \Leftrightarrow \begin{cases} \mathbf{preference}(\varphi_{t'}, \mathcal{I}_t^x) \\ \mathbf{inObs}(\varphi_{t'}, t) \\ \mathbf{focus}^{(-)}(\varphi_{t'}, d_j, \mathcal{I}_t^x, rev) \end{cases}$$

Exemple Pour illustrer ce biais, nous nous limitons ici à un problème de préférence entre deux informations et à un pas de temps :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\mathrm{alarm}_1, \mathrm{greenLight}_1\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \mathrm{alarm}_t \rightarrow \mathrm{outFuel}_t \\ R^b & \equiv & \mathrm{greenLight}_t \rightarrow \neg\, \mathrm{outFuel}_t \end{array} \right\} \\ \mathcal{D} & \equiv & \{\neg\, \mathrm{outFuel}_t\} \end{array}$$

Ici l'agent observe au pas de temps 1 une alarme et un voyant vert qui lui indique respectivement qu'il est en manque d'essence et qu'il ne l'est pas. Nous trouvons comme ignorances au pas de temps 1 :

$$\begin{array}{lcl} \mathcal{I}_1^a & = & \{\mathrm{alarm}_1\} \\ \mathcal{I}_1^b & = & \{\mathrm{greenLight}_1, \mathcal{D}(\neg\, \mathrm{outFuel}_1)\} \\ \mathcal{I}_1^c & = & \{R^a(1)\} \\ \mathcal{I}_1^d & = & \{R^b(2), \mathcal{D}(\neg\, \mathrm{outFuel}_1)\} \end{array}$$

Nous cherchons à déterminer si le littéral $\mathrm{greenLight}_1$ dans l'ensemble $\mathcal{I}_1^b$ est un biais de facilitation de préférence. L'ignorance est bien due à un problème de préférence du fait qu'aucune action ne rentre en jeu dans l'incohérence. De plus, un choix d'ignorance possible pour $\mathrm{greenLight}_1$ est d'ignorer une autre observation alarm_1 qui a pour alternative $\mathcal{I}_1^a$. Or dans l'état mental résultant de $\mathcal{I}_1^b$ le désir de ne pas être en manque d'essence n'est pas satisfait alors qu'il l'est dans l'état de croyance résultant de $\mathcal{I}_1^a$: $\mathrm{greenLight}_1$ correspond donc à un choix d'attrait vers un désir non satisfait. Nous avons donc toutes les conditions réunies pour définir l'ignorance de $\mathrm{greenLight}_1$ comme un biais de facilitation de préférence.

Facilitation de décision Pour le cas de la *facilitation de décision*, nous considérons que nous sommes dans une situation où :

(1) L'agent prend une mauvaise décision au pas de temps t, car il n'a pas observé $\varphi_{t'}$

(2) L'agent croit que son désir d_j n'est pas satisfait

(3) Il existe une autre observation ψ_i au pas de temps t qui mène à un $\neg d_j$

En d'autres termes, l'agent a pris une mauvaise décision en ne portant pas son attention sur $\varphi_{t'}$ car il existait une observation ψ_i qui lui indiquait un non-désir. De cette définition, nous en concluons que nous sommes face à un problème de décision ou de distorsion, car une observation $\varphi_{t'}$ a été ignorée à cause de l'action de l'agent. L'action peut donc être mauvaise en ne prenant pas en compte l'*inertie* (décision) ou en la prenant en compte (distorsion). De plus, du fait qu'il existe une autre observation ψ_i qui mène à un non-désir, cela veut dire que le littéral de désir d_j est ignoré dans l'ensemble où $\varphi_{t'}$ est ignoré car ψ_i est conservé et le désir d_j n'est pas satisfait. En d'autres termes, ignorer ce désir d_j montre que l'agent a concentré l'attention sur $\neg d_j$. Nous pouvons donc réutiliser l'opérateur $focus^{(-)}$ pour la définition de la facilitation de décision pour l'étape de décision ou de distorsion .Nous notons alors le biais $facilitation^{(D)}$ qui retourne vrai si

l'ignorance de $\varphi_{t'}$ dans l'ensemble $\mathcal{I}_t^x$ correspond à un biais de *facilitation de décision* pour le désir d_j :

$$\textbf{facilitation}^{(\mathbf{D})}(\varphi_{t'}, \mathcal{I}_t^x, d_j) \Leftrightarrow \left\{ \begin{array}{l} \textbf{decision}(\varphi_{t'}, \mathcal{I}_t^x) \text{ ou } \textbf{distorsion}(\varphi_{t'}, \mathcal{I}_t^x) \\ \textbf{inObs}(\varphi_{t'}, t) \\ d_j \in \mathcal{I}_x^t \\ \textbf{focus}^{(-)}(d_j, d_j, \mathcal{I}_t^x, diag) \\ \text{ou } \textbf{focus}^{(-)}(d_j, d_j, \mathcal{I}_t^x, dist) \end{array} \right.$$

Exemple Pour illustrer ce biais, nous nous limitons ici à pas de temps :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\text{alarm}_1, \neg\, \text{landPermission}_1\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \text{alarm}_t \rightarrow \text{outFuel}_t \\ R^b & \equiv & [\text{landPermission}_t]\, \text{land}_t \end{array} \right\} \\ \mathcal{T} & \equiv & \{\text{land}_1\} \\ \mathcal{D} & \equiv & \{\neg\, \text{outFuel}_t\} \end{array}$$

Nous avons ici un agent qui observe qu'il n'a pas la permission d'atterrir et une alarme lui indiquant qu'il est en manque d'essence. L'agent décide d'atterrir malgré qu'il n'ait pas la permission. Nous trouvons comme ignorances pour le pas de temps 1 :

$$\begin{array}{lcl} \mathcal{I}_1^a & = & \{\neg\, \text{landPermission}_1, \mathcal{D}(\neg\, \text{outFuel}_1)\} \\ \mathcal{I}_1^b & = & \{\neg\, \text{landPermission}_1, R_1^a\} \\ \mathcal{I}_1^c & = & \{\neg\, \text{landPermission}_1, \text{alarm}_1\} \\ \mathcal{I}_1^d & = & \{R_1^b, \mathcal{D}(\neg\, \text{outFuel}_1)\} \\ \mathcal{I}_1^e & = & \{R_1^b, R_1^a\} \\ \mathcal{I}_1^f & = & \{R_1^b, \text{alarm}_1\} \end{array}$$

Considérons que nous souhaitons déterminer si l'ignorance de $\neg\, \text{landPermission}_1$ dans $\mathcal{I}_1^a$ correspond à un biais de *facilitation de décision*. Cette ignorance est bien due à un problème de décision du fait que sans l'action land_1 l'ignorance de la permission d'atterrir n'est pas nécessaire. De plus, nous pouvons trouver que l'ignorance de l'autre proposition de $\mathcal{I}_1^a$ correspond à un choix d'attrait vers une perte, car nous pouvons ignorer l'observation de l'alarme à la place pour ne pas déduire le non-désir outFuel_1. Nous avons donc toutes les conditions réunies pour définir l'ignorance de $\neg\, \text{landPermission}_1$ comme un biais de *facilitation de décision*.

Enfin, nous notons l'opérateur *facilitation* qui retourne vrai si l'ignorance de $\varphi_{t'}$ est biais de facilitation, c'est-à dire un biais de facilitation de préférence ou de décision :

$$\textbf{facilitation}(\varphi_{t'}, \mathcal{I}_t^x, d_j) \Leftrightarrow \left\{ \begin{array}{l} \textbf{facilitation}^{(\mathbf{P})}(\varphi_{t'}, \mathcal{I}_t^x, d_j) \\ \text{ ou } \textbf{facilitation}^{(\mathbf{D})}(\varphi_{t'}, \mathcal{I}_t^x, d_j) \end{array} \right.$$

VII.2.2.b Biais de désengagement

Pour le biais de désengagement, nous considérons que le fait que l'attention vers un danger est difficile à modifier si l'attention est facilitée sur ce même danger

sur deux pas de temps à la suite :

(1) si ψ_i est ignoré pour faciliter l'attention sur le désir d_j au temps précédent
(2) si $\varphi_{t'}$ est ignoré pour faciliter l'attention sur le même désir d_j au temps suivant

La définition du désengagement est alors dépendante des biais de facilitation : nous devons trouver un biais de facilitation sur un même désir sur le pas de temps de l'ignorance et le pas de temps précédent. Nous notons alors le biais $disengagement$ qui retourne vrai si l'ignorance $\varphi_{t'}$ dans $\mathcal{I}_t^x$ correspond à un biais de désengagement :

$$\mathbf{disengagment}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \exists d_j \in \mathcal{D}, \mathbf{facilitation}(\varphi_{t'}, \mathcal{I}_t^x, d_j) \\ \exists \mathcal{I}_{t-1}^y \in S(\mathcal{I}_t^x) \\ \mathbf{facilitation}(\varphi_{t'}, \mathcal{I}_{t-1}^y, d_j) \\ \text{ ou } (j = t \text{ et } \mathbf{facilitation}(\varphi_{t'}, \mathcal{I}_{t-1}^y, d_{t-1})) \end{cases}$$

Exemple Pour illustrer ce biais, nous allons reprendre tout simplement les deux exemples pour le biais de facilitation sur deux pas de temps :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\mathrm{alarm}_1, \mathrm{greenLight}_1\} \\ Obs_2 & \equiv & \{\mathrm{alarm}_2, \neg\,\mathrm{landPermission}_2\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \mathrm{alarm}_t \rightarrow \mathrm{outFuel}_t \\ R^b & \equiv & \mathrm{greenLight}_t \rightarrow \neg\,\mathrm{outFuel}_t \\ R^c & \equiv & [\mathrm{landPermission}_t]\,\mathrm{land}_t \end{array} \right\} \\ \mathcal{T} & \equiv & \{\mathrm{land}_1, \mathrm{land}_2\} \\ \mathcal{D} & \equiv & \{\neg\,\mathrm{outFuel}_t\} \end{array}$$

Si l'agent décide d'ignorer $\neg\,\mathrm{landPermission}_2$ au pas de temps 2 alors, il tombe dans un biais de facilitation comme nous avons vu au VII.2.2.a. De plus, si l'agent décide d'ignorer au pas de temps 1 $\mathrm{greenLight}_1$ alors, il tombe là aussi dans un biais de facilitation (voir le premier exemple du biais de facilitation). Enfin, ces deux ignorances sont des biais de facilitation respectivement sur les désirs $\neg\,\mathrm{outFuel}_1$ et $\neg\,\mathrm{outFuel}_2$ qui sont le même désir par *inertie*. De ce fait, de telles ignorances correspondent à un biais de désengagement selon la définition ci-dessus.

VII.2.2.c Biais d'évitement

Pour le biais d'évitement, nous considérons que l'attention est dirigée vers une information opposée au danger si :

(1) l'agent ignore une observation $\varphi_{t'}$;
(2) $\varphi_{t'}$ implique un désir d_j non satisfait.

De cette définition, nous en concluons qu'il existe un autre choix d'ignorance pour $\varphi_{t'}$ qui consistait à croire que le désir d_j n'est pas satisfait. C'est-à-dire que l'agent a évité une perte, ce qui correspond à la définition de la caractéristique de l'aversion à la perte que nous avons abordé sous-section VII.1.4. Nous notons alors l'opérateur

$choiceAversion^{(-)}$ s'il existe un autre choix qui permettait de déduire $\neg d_j$:

$$\mathbf{choiceAversion}^{(-)}(\varphi_{t'}, d_j, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \exists \Psi \in choice(\varphi_{t'}, \mathcal{I}_t^x) \\ \mathbf{aversion}^{(-)}(d_j, B_t^x, B_t^{alt}) \end{cases}$$

avec $\mathcal{I}_t^{alt} = alt(\Psi, \mathcal{I}_t^x)$

Nous définissons alors le biais d'évitement par la fonction *avoid* qui retourne vraie si l'ignorance $\varphi_{t'}$ dans l'ensemble $\mathcal{I}_t^x$ correspond à un biais d'évitement :

$$\mathbf{avoid}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \mathbf{inObs}(\varphi_{t'}, t) \\ \exists d_j \in \mathcal{D}, \mathbf{choiceAversion}^{(-)}(\varphi_{t'}, d_j, \mathcal{I}_t^x) \end{cases}$$

Exemple Pour illustrer ce biais, nous reprenons exactement l'exemple du biais de facilitation de préférence (voir au VII.2.2.a). Si dans cet exemple, nous souhaitons déterminer si l'ignorance de alarm_1 dans $\mathcal{I}_1^a$ correspond à un biais d'évitement. Nous pouvons voir qu'un autre choix possible pour cette ignorance est d'ignorer $\mathrm{greenLight}_1$ et $\mathcal{D}(\neg\,\mathrm{outFuel}_1)$ et donc d'avoir un état de croyances alternatif où le désir de ne pas être en manque d'essence n'est pas satisfait. En ignorant l'alarme, l'agent a donc évité une perte et dirigé son attention sur d'autres informations.

VII.2.3 Optimisme

Le biais d'optimisme correspond au fait de croire à être moins exposé aux évènements négatifs Par exemple, un pilote observe des nuages et décide, malgré le danger, de continuer le trajet. Formellement :

(1) L'agent ignore une règle de raisonnement $\varphi_{t'}$ qui lui permettait de conclure à un désir d_j non satisfait.

De cette définition, nous en concluons qu'il existe un autre choix d'ignorance pour la règle $\varphi_{t'}$ qui consiste à croire que le désir d_j n'est pas satisfait. C'est-à-dire que l'agent a évité une perte, ce qui correspond à la définition de l'opérateur $choiceAversion^{(-)}$ que nous avons déjà défini pour le biais d'évitement. Nous définissons alors le biais d'optimisme par la fonction $optimism$ qui retourne vraie si l'ignorance de $\varphi_{t'}$ dans l'ensemble $\mathcal{I}_t^x$ correspond à un biais d'optimisme :

$$\mathbf{optimism}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \mathbf{isRule}(\varphi_{t'}) \\ \exists d_j \in \mathcal{D}, \mathbf{choiceAversion}^{(-)}(\varphi_{t'}, d_j, \mathcal{I}_t^x) \end{cases}$$

Exemple Pour illustrer ce biais, nous allons utiliser à nouveau l'exemple au VII.2.2.a du biais d'évitement. En effet, le biais d'optimisme a la même définition que le biais d'évitement, la seule différence est que c'est une règle de raisonnement qui est ignorée et non une observation. Donc en ignorant la règle R^a dans l'ensemble $\mathcal{I}_c^1$ nous trouvons aussi un autre choix qui consiste à ignorer $\mathrm{greenLight}_1$ et $\mathcal{D}(\neg\,\mathrm{outFuel}_1)$ et donc d'avoir un état de croyances alternatif où le désir de ne pas être en manque d'essence n'est pas satisfait. En ignorant l'alarme, l'agent a donc évité une perte et a décidé de croire qu'il ne sera pas en manque

d'essence dans son cas. L'optimisme est différent du biais d'évitement du fait que dans le cas du biais d'évitement, l'agent n'observe pas l'information qui implique le non désir là où dans le biais d'optimisme, l'agent observe l'information mais biaise son raisonnement pour croire qu'elle n'aura pas un impact négatif.

VII.2.4 Biais d'engagement

Le biais d'engagement correspond à la tendance à persister dans un comportement irrationnel malgré des résultats de plus en plus négatifs. En d'autres termes, nous voulons capturer le raisonnement de l'agent qui consiste à penser que si une action n'a pas marché, elle marchera la prochaine fois. D'un point de vue logique, nous considérons que l'agent est dans un biais d'engagement si :

(1) L'agent ignore les effets d'une action précédente à cause de l'incohérence entre ces effets et les observations du pas de temps t

(2) L'agent répète la même action au pas de temps t que celle du pas de temps $t-1$

De cette définition, nous en concluons que nous sommes face à un problème de préférence du fait que l'agent a préféré croire que l'action n'a pas eu l'effet attendu au vu des observations, mais décide de refaire la même action. Nous notons alors la fonction *commitment* qui retourne vraie si l'ignorance $\varphi_{t'}$ dans l'ensemble $\mathcal{I}_t^x$ correspond à un biais d'engagement :

$$\mathbf{commitment}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \mathbf{preference}(\varphi_{t'}, \mathcal{I}_t^x) \\ \mathbf{isConsequence}(\varphi_{t'}, t-1) \\ \mathbf{sameAction}(t, t-1) \end{cases}$$

Exemple Pour illustrer ce biais, considérons un exemple où nous nous limitons à un problème de préférence :

$$\begin{array}{lcl} Obs_1 & \equiv & \mathrm{alarm}_1 \\ Obs_2 & \equiv & \mathrm{alarm}_2 \\ \mathcal{R} & \equiv & \begin{cases} R^a & \equiv & \mathrm{alarm}_t \rightarrow \mathrm{stall}_t \\ R^b & \equiv & (\mathrm{push}_t \wedge \mathrm{stall}_t) :: \neg\, \mathrm{stall}_{t+1} \end{cases} \\ \mathcal{T} & \equiv & \{\mathrm{push}_1, \mathrm{push}_2\} \end{array}$$

Dans cet exemple, l'agent peut observer une alarme qui lui indique un décrochage et décide par deux fois de pousser le manche. L'agent croit que cette action a pour effet de sortir de la situation de décrochage. Au pas de temps 2, nous trouvons comme ignorances :

$$\begin{array}{lcl} \mathcal{I}_2^a & = & \{R_1^a\} \\ \mathcal{I}_2^b & = & \{R_2^a\} \\ \mathcal{I}_2^c & = & \{R_1^b\} \\ \mathcal{I}_2^d & = & \{\mathrm{alarm}_1, keep_{\mathrm{alarm}}^{(val)}(2)\} \\ \mathcal{I}_2^e & = & \{\mathrm{alarm}_2, keep_{\mathrm{alarm}}^{(val)}(2)\} \end{array}$$

Ici l'ignorance de R_1^b dans $\mathcal{I}_1^c$ correspond à un biais d'engagement du fait que la règle R_1^b correspond à l'ignorance des effets de l'action précédente, de plus, la même action push est répétée sur le temps 1 et 2. L'agent s'engage donc dans l'action de pousser le manche de contrôle en pensant que c'est la meilleure action pour sortir du décrochage.

VII.2.5 Illusion de contrôle

L'illusion de contrôle est la tendance à croire que nous avons une influence sur des événements indépendants à notre contrôle. Par exemple, un agent croit que tirer sur le manche de contrôle a eu pour effet de sortir du décrochage et décide de tirer à nouveau sur le manche quand il se retrouve à nouveau en décrochage. Nous considérons alors la définition suivante pour l'illusion de contrôle :

(1) Une règle de raisonnement $\varphi_{t'}$ est ignorée à cause d'une décision incohérente
(2) L'action effectuée au temps t a déjà été effectuée par le passé
(3) Cette action avait mené à un changement d'état d'un désir d_j qui est passé de faux à vrai
(4) Ce même désir d_j n'est pas satisfait au temps t

De cette définition, nous en concluons que c'est un problème de distorsion ou de décision du fait que c'est l'action effectuée qui est la cause de l'ignorance de $\varphi_{t'}$. Nous devons trouver une action qui a par le passé provoqué un changement positif que nous avons défini sous-section VII.1.5. Nous définissons l'opérateur $successAction$ qui retourne vrai si l'action entre l'état de croyances B_{t+1} B_t a eu pour effet un changement positif du désir d_j :

$$\mathbf{successAction}(d_j, B_{t+1}, B_t) \Leftrightarrow \begin{cases} \mathbf{changeRevision}^{(+)}(d_j, B_{t+1}, B_t) \\ \text{ou } \mathbf{changeUpdate}^{(+)}(d_j, B_{t+1}, B_t) \end{cases}$$

Nous devons vérifier ensuite que cette action qui a pour effet un changement positif sur un désir est la même action effectuée qu'au pas de temps t. Pour cela, nous utilisons la caractéristique $sameAction$ définie sous-section VII.1.5. Enfin, nous vérifions que le changement est fait sur un désir qui n'est pas satisfait au pas de temps t. Nous utilisons pour cela la caractéristique $sameDesire$. Nous définissons alors le biais d'illusion de contrôle par la fonction $controlIllusion$ qui retourne vraie si l'ignorance de $\varphi_{t'}$ dans l'ensemble $\mathcal{I}_t^x$ est une illusion de contrôle :

$$\mathbf{controlIllusion}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \exists \mathcal{I}_i^y \in S(\mathcal{I}_t^x), i < t \\ \exists d_j \in \mathcal{D} \\ \mathbf{decision}(\varphi_{t'}, \mathcal{I}_t^x) \vee \mathbf{distorsion}(\varphi_{t'}, \mathcal{I}_t^x) \\ \mathbf{isRule}(\varphi_{t'}) \\ \mathbf{notSatisfied}(d_j, B_t^x) \\ \mathbf{successAction}(d_j, B_t^x, B_i^y) \\ \mathbf{sameAction}(B_t^x, B_i^y) \\ \mathbf{sameDesire}^{(-)}(d_j, B_t^x, B_i^y) \end{cases}$$

Exemple Pour illustrer ce biais, nous allons considérer un exemple sur trois pas de temps :

$$
\begin{array}{lcl}
Obs_1 & \equiv & \{\mathrm{alarm}_1\} \\
Obs_2 & \equiv & \{\neg\,\mathrm{alarm}_2\} \\
Obs_3 & \equiv & \{\mathrm{alarm}_3\} \\
\mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \mathrm{alarm}_t \rightarrow \mathrm{stall}_t \\ R^b & \equiv & \neg\,\mathrm{alarm}_t \rightarrow \neg\,\mathrm{stall}_t \\ R^c & \equiv & [\neg\,\mathrm{stall}_t]\,\mathrm{pull}_t \end{array} \right. \\
\mathcal{D} & \equiv & \neg\,\mathrm{stall}_t \\
\mathcal{T} & \equiv & \{\mathrm{pull}_1, \mathrm{push}_2, \mathrm{pull}_3\}
\end{array}
$$

Ici l'agent peut observer au premier pas de temps une alarme qui lui indique un décrochage et décide de tirer le manche de contrôle, ce qui n'est pas cohérent du fait que tirer le manche nécessite de ne pas être en décrochage. Au deuxième pas de temps, l'agent observe que l'alarme est désactivée, ce qui lui indique qu'il n'est plus en décrochage. L'agent décide de pousser le manche. Enfin, au troisième pas de temps, l'agent peut observer de nouveau l'alarme, ce qui indique un décrochage. L'agent décide une nouvelle fois de tirer le manche.

Considérons que nous sommes dans le scénario où au pas de temps 1 l'agent ignore R_1^c et $\mathcal{D}(\neg\,\mathrm{stall}_1)$ pour effectuer l'action pull et au pas de temps 2 ignore $keep_2^{(val)}(\mathrm{alarm})$ pour mettre à jour le fait que l'alarme se désactive. Dans un tel scénario, nous trouvons comme ignorances au pas de temps 3 :

$$
\begin{array}{lcl}
\mathcal{I}_3^a & = & \{keep_3^{(val)}(\mathrm{alarm}), \mathrm{alarm}_3\} \\
\mathcal{I}_3^b & = & \{keep_3^{(val)}(\mathrm{alarm}), R^a(3)\} \\
\mathcal{I}_3^c & = & \{keep_3^{(val)}(\mathrm{alarm}), R^c(3), \mathcal{D}(\neg\,\mathrm{stall}_3)\}
\end{array}
$$

Nous souhaitons déterminer si l'ignorance de R^c dans l'ensemble $\mathcal{I}_3^c$ est une illusion de contrôle. Tout d'abord, l'ignorance est bien due à un problème de distorsion du fait que c'est l'action et la prise en compte de l'inertie qui est responsable de l'incohérence. De plus, le désir de ne pas être en décrochage n'est pas satisfait dans l'état de croyances résultant du MCS M_3^c. Enfin, nous pouvons trouver que l'action de tirer au pas de temps 1 a eu pour effet que le désir de ne pas être en décrochage est passé de non satisfait à satisfait du fait que l'alarme se désactive au pas de temps 2. Or l'action de tirer est répétée au pas de temps 3 alors que le désir de ne pas être en décrochage n'est pas satisfait. L'ignorance de la règle R^c dans $\mathcal{I}_3^c$ respecte donc toutes les conditions pour être un biais d'illusion de contrôle.

VII.2.6 Faux souvenirs

Les faux souvenirs correspondent au fait de se souvenir d'une information qui n'existe pas ou s'en souvenir de manière erronée. Il a été montré que ces faux souvenirs étaient favorisés lorsqu'un agent portait son attention sur des informations centrales à leurs objectifs [Kaplan et al., 2016]. Nous considérons alors la définition suivante pour les faux souvenirs :

(1) L'agent prend une mauvaise décision au pas de temps t, car il ne se souvenait pas de $\varphi_{t'}$ ou se souvenait de $\neg\varphi_{t'}$ au lieu de $\varphi_{t'}$.
(2) L'agent croit que son désir d_j n'est pas satisfait.
(3) Il existe une autre observation ψ_i au pas de temps t qui mène à $\neg d_j$

De cette définition, nous en concluons qu'un faux souvenirs correspond à ignorer une proposition d'*inertie* ou une croyance à propos du passé. En effet, se souvenir de manière erronée d'une information $\varphi_{t'}$ peut-être dû au fait que l'*inertie* des croyances ait changé de manière inattendue (l'agent croit que $\varphi_{t'}$ a changé entre deux pas de temps sans aucune raison). Un souvenir erroné peut être aussi dû à une ignorance d'une croyance à propos du passé. Par exemple, si j'ignore la croyance au pas de temps 2 qu'il y a une alarme au temps 0 alors par inertie, je ne vais pas croire en l'alarme au pas de temps 2. De plus, nous pouvons observer que la définition du faux souvenir est très proche de la définition de la facilitation de décision abordée au VII.2.2.a. L'agent concentre son attention sur une information qui mène vers un danger et ignore alors une information. Nous allons donc utiliser l'opérateur $focus^{(-)}$ pour définir ce biais. Nous notons la fonction $falseMemory$ qui retourne vraie si pour l'ignorance $\varphi_{t'}$ dans l'ensemble $\mathcal{I}_t^x$ correspond à un faux souvenir :

$$\mathbf{falseMemory}(\varphi_{t'}, \mathcal{I}_t^x) \Leftrightarrow \begin{cases} \mathbf{distortion}(\varphi_{t'}, \mathcal{I}_t^x) \\ \exists d_j \in \mathcal{D}, \exists \psi_i \in \mathcal{I}_t^x, \psi_i \neq \varphi_{t'} \\ \mathbf{focus}^{(-)}(\psi_i, d_j, \mathcal{I}_t^x) \end{cases}$$

Exemple Pour illustrer ce biais, nous considérons l'exemple suivant :

$$\begin{array}{lcl} \mathcal{I}nit & \equiv & \{\neg\,\mathrm{landPermission}_0\} \\ Obs_1 & \equiv & \{\mathrm{alarm}_1\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R^a & \equiv & \mathrm{alarm}_t \rightarrow \mathrm{outFuel}_t \\ R^b & \equiv & [\mathrm{landPermission}_t]\,\mathrm{land}_t \end{array} \right\} \\ \mathcal{T} & \equiv & \{\mathrm{land}_1\} \\ \mathcal{D} & \equiv & \{\neg\,\mathrm{outFuel}_t\} \end{array}$$

Dans cet exemple, l'agent croit au pas de temps 0 qu'il n'a pas le droit d'atterrir. Au pas de temps 1, il observe une alarme qui lui indique un manque d'essence et

décide d'atterrir. Nous trouvons comme ignorance au pas de temps 1 :

$$
\begin{array}{rcl}
\mathcal{I}_1^a & = & \{\neg\,\mathrm{landPermission}_0, \mathrm{alarm}_1\} \\
\mathcal{I}_1^b & = & \{keep_1^{(val)}(\mathrm{landPermission}), \mathrm{alarm}_1\} \\
\mathcal{I}_1^c & = & \{keep_1^{(known)}(\mathrm{landPermission}), \mathrm{alarm}_1\} \\
\mathcal{I}_1^d & = & \{\neg\,\mathrm{landPermission}_0, R^a(1)\} \\
\mathcal{I}_1^e & = & \{keep_1^{(val)}(\mathrm{landPermission}), R^a(1)\} \\
\mathcal{I}_1^f & = & \{keep_1^{(known)}(\mathrm{landPermission}), R^a(1)\} \\
\mathcal{I}_1^g & = & \{R^b(1), \mathrm{landPermission}_1\} \\
\mathcal{I}_1^h & = & \{R^b(1), R^a(1)\} \\
\mathcal{I}_1^i & = & \{R^b(1), \mathcal{D}(\neg\,\mathrm{outFuel}_1)\} \\
\mathcal{I}_1^j & = & \{\neg\,\mathrm{landPermission}_0, \mathcal{D}(\neg\,\mathrm{outFuel}_1)\} \\
\mathcal{I}_1^k & = & \{keep_1^{(val)}(\mathrm{landPermission}), \mathcal{D}(\neg\,\mathrm{outFuel}_1)\} \\
\mathcal{I}_1^l & = & \{keep_1^{(known)}(\mathrm{landPermission}), \mathcal{D}(\neg\,\mathrm{outFuel}_1)\}
\end{array}
$$

Considérons que nous souhaitons savoir si l'ignorance de $keep_1^{(val)}(\mathrm{landPermission})$ dans l'ensemble $\mathcal{I}_1^k$ équivaut à un faux souvenir. Nous pouvons voir que dans l'ensemble $\mathcal{I}_1^k$, il existe une autre ignorance $\mathcal{D}(\neg\,\mathrm{outFuel}_1)$ tel que l'opérateur $focus^{(-)}$ est vrai pour cette ignorance. L'attention est portée sur l'alarme qui implique que le désir de ne pas être en manque d'essence n'est pas satisfait. De ce fait, l'ignorance de $keep_1^{(val)}(\mathrm{landPermission})$ regroupe toutes les conditions pour être un faux souvenir.

VII.2.7 Conclusion

Nous avons dans cette section pu montrer que nous pouvions exprimer différents biais cognitifs à partir des caractéristiques que nous avons définies section VII.1. Nous arrivons à la fois à exprimer des biais liés aux émotions (optimisme, biais d'attention, faux souvenir, illusion de contrôle), à la mémoire (faux souvenir), au coût cognitif d'une révision (biais de confirmation) ou à une décision irrationnelle (biais d'engagement). Notre but dans cette section n'étant pas de faire une liste exhaustive de tous les biais pouvant être exprimés dans notre modèle, nous nous sommes concentrés sur les définitions des biais les plus représentatifs capturés par notre modèle afin que le lecteur puisse comprendre l'idée de l'expression des biais sur quelques exemples.

Toutefois, bien que notre modèle soit assez expressif pour représenter de nombreux biais, il ne l'est pas assez pour capturer l'ensemble des biais cognitifs connus dans la littérature. Par exemple, notre modèle se base sur une logique de croyance mono-agent et non multi-agent, par conséquent tous les biais sociaux ne sont pas capturés. Nous reviendrons sur la limitation de l'expressivité de notre modèle dans les sections suivantes.

Maintenant que nous pouvons déterminer si une ignorance correspond à un biais ou non, nous devons déterminer quels sont les états de croyances les plus plausibles.

Pour cela, comme introduit chapitre IV nous allons utiliser un algorithme de filtrage basé sur la recherche de ces biais. Nous présentons un premier algorithme naïf dans la prochaine section.

VII.3 Un premier algorithme de filtrage

Nous avons pu définir de manière formelle un ensemble de biais cognitif pour expliquer les raisons d'une ignorance. Nous devons maintenant comparer les différents états de croyances résultant de ces ignorances pour déterminer les états les plus plausibles qui représentent l'explication la plus probable pour l'accident modélisé. Pour cela, nous allons définir un score de plausibilité sur les ignorances et à partir de ce score définir la relation de plausibilité permettant de déterminer l'état de croyance le plus plausible à un instant t.

VII.3.1 Score de plausibilité

Un état de croyance à un instant t (*i.e* B_t^x) peut être décrit par un ensemble d'ignorances $\mathcal{I}_t^x$ comme nous avons vu équation VI.5, c'est-à-dire à partir d'un chemin de correction x. Le score de plausibilité que nous cherchons à déterminer doit pouvoir représenter à quel degré l'ensemble des ignorances de $\mathcal{I}_t^x$ sont une explication possible du comportement de l'agent. Pour cela, nous allons considérer que plus $\mathcal{I}_t^x$ contient des ignorances qui peuvent être expliquées cognitivement, plus son score de plausibilité est haut. Nous allons, dans un premier temps, définir ce que nous entendons par explication cognitive puis définir ce score de plausibilité.

VII.3.1.a Définition d'une explication cognitive

Pour déterminer si une ignorance a une explication possible, nous devons différencier deux types d'explications que nous englobons sous le terme d'explication cognitive :

(1) les explications dues à un biais cognitif ;

(2) les explications dues à un comportement attendu de l'agent.

Le premier type d'explication correspond aux définitions des biais section VII.2 : l'agent ignore une croyance, règle ou observation qui peut être expliquée par un biais que nous pouvons retrouver dans la littérature des sciences humaines.

Le deuxième type d'explication correspond à des ignorances qui sont des changements de croyances attendues dans un agent rationnel. Nous considérons trois changements attendus :

— L'*extrapolation* car le monde a changé et les croyances doivent être mises à jour.

— La *mise à jour* car les actions de l'agent ont changé le monde et les croyances doivent être mises à jour.

— Un désir non satisfait par l'agent qui est déduit par les croyances de l'agent. C'est-à-dire que l'agent, par ses observations, peut déduire qu'il n'est pas dans un état satisfaisant son désir. Dans un tel cas, l'agent ignore son désir et nous considérons que cette ignorance est attendue.

Du fait que ces ignorances sont attendues, nous considérons que ce sont des explications suffisantes en elles-mêmes. Or, nous avons vu section VII.1 que nous pouvons déterminer si une ignorance est due à l'*extrapolation* ou à la *mise à jour* par les caractéristiques $\mathbf{extrapolation}(\varphi_{t'}, \mathcal{I}_t^x)$ et $\mathbf{update}(\varphi_{t'}, \mathcal{I}_t^x)$ (sous-section VII.1.3). Enfin, nous pouvons déterminer si l'ignorance d'un désir représente le fait que l'agent croit qu'il est dans un état qui ne satisfait pas un désir grâce à la caractéristique $\mathbf{notSatisfied}(d_{t'}, B_t^x)$. Ainsi, toute ignorance ayant l'une de ces trois caractéristiques est expliquée par défaut du fait que l'ignorance correspond à un changement attendu dans les croyances de l'agent.

De cette distinction entre les deux types d'explications, nous en concluons qu'il est nécessaire de déterminer un score de plausibilité que pour un ensemble d'ignorances qui ne sont pas dues à l'*extrapolation*, la *mise à jour* et un désir non satisfait par l'agent du fait que celles-ci ont une explication par défaut.

VII.3.1.b Définition du score de plausibilité

Pour définir à quel degré un ensemble d'ignorance $\mathcal{I}_t^x$ correspond à une explication plausible, nous allons tout simplement déterminer le pourcentage d'ignorance dans $\mathcal{I}_t^x$ qui ont une explication possible. C'est-à-dire trouver le nombre d'ignorances ayant au moins une explication plausible divisé par le nombre total d'ignorances de $\mathcal{I}_t^x$. Nous notons $\mathcal{B}ias = \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ l'ensemble des fonctions booléennes représentant les biais cognitifs. Nous notons alors $Pl(\mathcal{I}_t^x)$ le score de plausibilité de l'ensemble des ignorances $\mathcal{I}_t^x$. Le score $Pl(\mathcal{I}_t^x)$ est retourné par l'algorithme 2. L'algorithme consiste à : pour toutes les ignorances $\varphi_{t'}$ dans $\mathcal{I}_t^x$ qui ne sont pas dues à l'*extrapolation* ou à la *mise à jour*, ajouter 1 au nombre d'explications trouvé lorsque qu'au moins une fonction booléenne de biais retourne vraie. Le score total est alors le nombre d'explications trouvé rapporté au total du nombre d'ignorances qui ne sont pas des *extrapolations* ou *mise à jour* dans $\mathcal{I}_t^x$.

VII.3.2 Définition de la relation de plausibilité

Nous avons pu définir dans la section précédente un score de plausibilité pour un ensemble d'ignorances. Nous devons maintenant donc comparer les ensembles d'ignorances pour déterminer l'état de croyance le plus plausible. Nous allons donc définir une relation de plausibilité entre les $\mathcal{I}_t^x$ d'un même pas de temps dans les différents scénarios. Pour cela, nous considérons que plus la plausibilité Pl d'un ensemble $\mathcal{I}_t^x$ est haut, plus l'état de croyances résultant de ces ignorances est plausible. Nous notons alors la relation d'ordre $\leq_{Pl}$ qui permet d'exprimer qu'un état de croyance a une plausibilité plus haut qu'un autre état de croyance si le

Algorithm 2 Retourne le score de plausibilité Pl de $\mathcal{I}_t^x$

$nbrExplain = 0$
$total = 0$
for $\varphi_{t'}$ **in** $\mathcal{I}_t^x$ **do**
  if not $(\mathbf{extrapolation}(\varphi_{t'}, \mathcal{I}_t^x)$ **or** $\mathbf{update}(\varphi_{t'}, \mathcal{I}_t^x))$ **then**
    $total+ = 1$
    $j = 0$
    $explain = false$
    while $j < |\mathcal{B}ias|$ **and not** $explain$ **do**
      $\mathcal{F} = \mathcal{B}ias(j)$
      if $\mathcal{F}(\varphi_{t'}, \mathcal{I}_t^x)$ **then**
        $nbrExplain+ = 1$
        $explain = true$
      $j+ = 1$
return $\frac{nbrExplain}{total}$

score de plausibilité de $\mathcal{I}_t^x$ correspond Pl est lui aussi plus haut :

$$B_t^y \leq_{Pl} B_t^x \quad \Leftrightarrow \quad Pl(\mathcal{I}_t^y) \leq Pl(\mathcal{I}_t^x)$$

Nous notons alors $\mathcal{K}$ le seuil de plausibilité (*threshold*) pour lequel tout état de croyances ayant un score de plausibilité inférieur à $\mathcal{K}$ est considéré comme pas assez plausible pour être une explication de l'accident. Par exemple, avec $\mathcal{K} = 0.8$, seuls les scénarios dont les états de croyances sont expliqués au moins à 80% sont considérés. Ce seuil permet de laisser une liberté à l'utilisateur du modèle d'explorer à sa guise seulement les scénarios avec une explication fournit par notre modèle ou des scénarios dont il n'existe pas d'explication avec les biais définis. L'utilisateur peut alors potentiellement repérer dans ces scénarios des motifs pouvant être exprimés avec la taxonomie formelle du modèle et être ajoutés comme une explication possible.

VII.3.3 Conclusion

Nous avons pu définir dans cette section le score de plausibilité pour un ensemble d'ignorances qui permet d'évaluer le degré d'explication de celui-ci en fonction des biais définis à partir de notre taxonomie. Ce score sert ensuite à définir une relation d'ordre entre les états mentaux pour déterminer les états les plus plausibles. Nous donnons une liberté pour l'utilisateur sur le seuil de plausibilité pour filtrer les scénarios les plus plausibles afin que l'utilisateur puisse explorer des scénarios sans explication connues pour le modèle afin de potentiellement pouvoir définir d'autres biais cognitifs.

Toutefois, notre calcul de score de plausibilité est encore trop simple. Nous ne prenons pas en compte par exemple l'enchaînement des biais dans le scénario qui

peut renseigner sur la plausibilité du scénario. Par exemple, un biais d'attention sur le même désir deux fois de suite est plus plausible qu'un biais d'attention sur un désir suivi d'un biais d'attention sur la contradiction de ce même désir (*e.g* un biais de facilitation sur le danger puis un biais d'évitement du danger). D'autres limites sur ce premier algorithme naïf sont à prendre en compte et nous les aborderons plus en détails dans les sections suivantes.

Cependant, nous pouvons déjà appliqué ce filtrage par plausibilité sur les différents cas d'études que nous avons présentées le long de cette thèse : l'accident du vol 447 (section IV.1) et du vol 5148 (section VI.1). La section suivante discutera du filtrage réalisé sur ces deux cas d'études.

VII.4 Application du modèle d'évaluation

VII.4.1 Accident du vol 447

Nous allons dans cette section discuter des résultats pour un filtrage sur l'accident 447 avec un seuil à $\mathcal{K} = 1$, c'est-à-dire que toutes les ignorances doivent être expliquées pour que le scénario soit considéré comme plausible.

Tout d'abord, sans filtrage (*i.e* $\mathcal{K} = 0$) nous trouvons 907 935 scénarios. C'est beaucoup plus que les 903 scénarios trouvés par l'application de notre première version du calcul de diagnostic sans itération et sans prise en compte de l'inertie. Cette explosion de scénario s'explique par le fait qu'en augmentant l'expressivité du modèle en prenant en compte l'inertie, nous augmentons les corrections possibles et donc états de croyances possibles. De plus, en utilisant un algorithme par itération, nous avons pu voir section VI.7 que cette méthode retournait plus de diagnostics que les simples diagnostics minimaux. Par conséquent, cela augmente aussi les états de croyances possibles. Toutefois, avec $\mathcal{K} = 1$, nous trouvons 3825 scénarios. Le filtrage permet donc de réduire considérablement les scénarios plausibles. Nous allons discuter pour chaque pas de temps des différents biais qui permettent d'expliquer les ignorances et qui permettent de filtrer les scénarios. Pour une faciliter de lecture, nous omettons dans les exemples d'ignorances, les propositions d'*inertie* qui sont dues à une *mise à jour* ou à une *extrapolation* du fait qu'elles sont considérées comme explicable par défaut.

VII.4.1.a Biais du premier pas de temps

Au premier pas de temps, une majorité des ignorances peuvent être expliquées par le biais de facilitation, d'optimisme et d'évitement. En effet, l'agent se retrouve dans une situation où des informations sont en contradiction entre elles et avec des désirs de l'agent. L'agent doit alors préférer une information qui l'amène de toute façon vers un désir non satisfait (*i.e* être en décrochage ou en survitesse).

Par exemple, si nous prenons l'ignorance :

$$\mathcal{I}_1^e \quad = \quad \{R^b(1), \mathrm{alarm}_1, \mathcal{D}(\neg\, \mathrm{overspeed}_1)\}$$

L'ignorance de alarm_1 est à la fois un biais de facilitation et d'évitement. En effet, cette ignorance a un autre choix possible, c'est-à-dire qu'il existait un autre chemin de correction (voir section VI.6) qui est $\Psi = \{\mathrm{acceleration}_1, R^g(1), \mathcal{D}(\neg\, \mathrm{stall}_1)\}$ où l'agent préfère ignorer l'accélération à la place de l'alarme. Si l'agent fait ce choix, il se retrouve dans un état où son désir de ne pas être en décrochage n'est pas satisfait. En ignorant l'alarme, il évite que ce désir de ne pas être en décrochage ne soit pas satisfait et respecte la définition du biais d'évitement. Enfin, s'il avait fait le choix d'ignorer l'accélération à la place de l'alarme son désir de ne pas être en survitesse serait satisfait. De ce fait, l'agent s'est concentré sur l'accélération qui lui indique un danger de survitesse. Nous retrouvons la définition formelle du biais de facilitation. De plus, l'ignorance du désir $\neg\, \mathrm{overspeed}_1$ est considérée comme attendu du fait que l'agent déduit qu'il est en survitesse par son état de croyances.

Nous trouvons que l'ignorance $R^b(1)$ a un même choix possible avec $\Psi = \{\mathrm{acceleration}_1, R^g(1), \mathcal{D}(\neg\, \mathrm{stall}_1)\}$. Par conséquent, nous pouvons déduire que là aussi l'agent a évité que son désir de ne pas être en décrochage ne soit pas satisfait. En ignorant une règle de raisonnement, l'agent tombe sur la définition du biais d'optimisme.

Toutefois, certaines ignorances sont filtrées comme :

$$\{\mathrm{acceleration}_1, \mathcal{D}(\neg\, \mathrm{stall}_1), \mathrm{alarm}_1, \mathrm{buffet}_1\}$$

Ici les observations de l'alarme et du buffet sont ignorés et l'opérateur $focus^{(-)}$ retourne que l'agent s'est concentré sur l'alarme et le buffet pour ignorer l'accélération. Toutefois, ces observations étant ignorées, l'agent ne peu pas se concentrer dessus. La définition des biais d'attention n'est alors pas respecté. Cet ensemble d'ignorances n'est alors pas gardé comme un état de croyance plausible.

VII.4.1.b Biais du deuxième pas de temps

Au deuxième pas de temps, nous trouvons qu'une majorité des ignorances ne dépassent pas le seuil $\mathcal{K} = 1$, permettant ainsi de filtrer de nombreux scénarios. En effet, de nombreux états contiennent l'ignorance du désir de ne pas être en décrochage, par exemple :

$$\mathcal{I}_2^o \quad = \quad \{\mathcal{D}(\neg\, \mathrm{stall}_2), R^g(2)\}$$

Le problème est que si au temps 1 l'agent ne peut pas déduire qu'il est en décrochage, alors il ne peut pas déduire qu'il est en décrochage au temps 2. Aucune observation lui permet de déduire cette information. En d'autres termes, le changement de valeur de vérité sur le désir n'est pas attendu, il n'existe alors pas d'explication de cette ignorance.

Nous trouvons toutefois plusieurs ignorances qui peuvent être pleinement expliquées. Par exemple :

$$\mathcal{I}_2^p \quad = \quad \{R^e(1), \mathcal{D}(\neg \, \mathrm{overspeed}_2)\}$$

Si l'état de croyances précédent de l'agent permettait de déduire qu'il n'était pas en décrochage, ces ignorances peuvent être expliquées. Ce MCS correspond à la définition du biais d'engagement du fait que l'agent ignore les effets de l'action précédente tout en répétant l'action précédente. Le désir $\neg \, \mathrm{overspeed}_2$ est déduit par l'agent du fait qu'il croit qu'il n'est pas en décrochage et que sa vitesse verticale augmente. Enfin, nous avons les ignorances :

$$\begin{array}{rcl} \mathcal{I}_2^q & = & \{\mathrm{VS} \uparrow_2\} \\ \mathcal{I}_2^r & = & \{R^d(2)\} \end{array}$$

qui représentent respectivement un biais d'évitement et d'optimisme du fait que l'agent évite de conclure qu'il est en survitesse en ignorant l'observation de la vitesse verticale qui augmente ou en ignorant la règle de raisonnement qu'à partir de la vitesse verticale cela lui permet de conclure qu'il est en survitesse.

VII.4.1.c Biais du troisième pas de temps

Au troisième pas de temps, il existe peu d'ignorances permettant d'atteindre le seuil $\mathcal{K} = 1$. Dans la même logique qu'au deuxième pas de temps certains états de croyance contiennent des ignorances de désir qui ne peuvent être déduites par l'agent. Si l'agent ignore son désir de ne pas être en décrochage alors que rien ne lui permet de déduire qu'il est en décrochage. Par exemple, l'ensemble d'ignorances :

$$\mathcal{I}_3^s \quad = \quad \{\mathrm{FDpull}_3, \mathcal{D}(\neg \, \mathrm{stall}_3)\}$$

Ici, si l'état de croyances à $t = 2$ ne permet pas de déduire que l'agent est en décrochage, alors l'ignorance du désir n'est pas attendue. Par conséquent, l'ensemble d'ignorance est filtré, car il n'est pas pleinement expliqué.

Si nous prenons le scénario qui contient $\mathcal{I}_2^p$, nous pouvons trouver comme MCS :

$$\mathcal{I}_3^t \quad = \quad \{R^e(2), \mathcal{D}(\neg \, \mathrm{overspeed}_3), \mathrm{FDpull}_3, \mathrm{FDpull}_2, R^h(3)\}$$

L'ignorance de la règle R^e correspond à un biais de confirmation, car si l'agent choisit de garder la règle, il doit ignorer l'observation de la vitesse verticale du temps 3 et sa croyance sur la vitesse verticale au temps 2. L'ignorance de FDpull_3 est un biais d'attention, car l'agent a porté son attention sur la vitesse verticale qui l'a mené à croire que son désir de ne pas être en survitesse n'est pas satisfait. L'ignorance de FDpull_2 est un faux souvenir du fait que l'attention est portée sur la vitesse verticale. Enfin, la règle R^h est un biais d'optimisme, car si l'agent

garde la règle R^h il doit croire qu'il n'est pas en survitesse, ce qui par conséquence l'oblige à croire qu'il est en décrochage. Il évite donc de croire que son désir de ne pas être en décrochage n'est pas satisfait en faisant l'action push_3. Ce MCS étant pleinement expliqué, il n'est pas filtré.

VII.4.1.d Biais du quatrième pas de temps

Au quatrième pas de temps, nous tombons dans un cas similaire que le troisième pas de temps où une majorité des ignorances est en dessous du seuil $\mathcal{K} = 1$. La majorité des ignorances remettent en cause les croyances du passé sans pour autant que ce soit un faux souvenir. Par exemple l'ignorance :

$$\mathcal{I}_4^u \quad = \quad \{\mathrm{alarm}_4, keep_{VS\uparrow}^{(val)}(4), \mathrm{VS}\uparrow_4\}$$

Ici l'ignorance de $keep_{VS\uparrow}^{(val)}(4)$ n'est pas expliquée car l'agent se retrouve dans un état où aucun de ses désirs n'est satisfait. Par conséquent, cette ignorance ne peut être expliquée par un faux souvenir du fait que l'agent ne s'est pas concentré sur une information en particulier.

$$\mathcal{I}_4^v \quad = \quad \{\mathrm{alarm}_4, \mathcal{D}(\neg\,\mathrm{overspeed}_4)\}$$

Nous trouvons comme choix possible pour l'ignorance de l'alarme : $\{\mathrm{FDpull}_4, \mathcal{D}(\neg\,\mathrm{stall}_4)\}$. Nous avons donc l'ignorance de l'alarme qui permet d'éviter d'ignorer le directeur de vol et un désir qui était satisfait dans l'état précédent. Par conséquent, le directeur de vol confirme une croyance précédente. Ignorer l'alarme revient à faire un biais de confirmation.

VII.4.1.e Conclusion sur le vol 447

Au vu du nombre de scénarios possible trouvé après filtrage pour l'accident du Rio-Paris, nous avons présenté ici un sous-ensemble de scénarios qui nous semble assez représentatif de l'ensemble des scénarios trouvés. Les autres scénarios étant des variations dans les propositions ignorées par rapport à ceux présenté mais sans pour autant changer forcément l'explication par les biais.

Nous pouvons voir que nous trouvons des scénarios conformes aux explications trouvées par le BEA :

- Le pilote porte son attention sur la vitesse qui augmente par un biais de facilitation en omettant l'alarme.
- Le pilote tombe dans un biais de confirmation en considérant que l'alarme n'est pas pertinente du fait des indications du directeur de vol.

Toutefois nous trouvons d'autres explications comme :

- Un biais d'évitement sur la situation de décrochage. C'est-à-dire que le pilote a tellement peur d'être en décrochage qu'il porte son attention sur toute information n'indiquant pas un décrochage.

— Un biais d'engagement par le fait que l'agent continue à tirer le manche de contrôle malgré que la situation ne change pas.
— Un biais d'optimisme sur par exemple les vibrations. C'est-à-dire que le pilote considère que dans sa situation les vibrations n'indique sûrement pas un décrochage et qu'il n'y a donc pas raison de s'inquiéter d'un décrochage.
— un faux souvenir sur les indications du directeur de vol. C'est-à-dire que le pilote oublie que le directeur de vol lui indiquait de tirer le manche en concentrant son attention sur la vitesse verticale.

Ainsi, notre modèle d'*évaluation* au-delà de retrouver les mêmes explications que l'enquête du BEA trouve d'autres explications cohérentes avec les décisions du pilote. De ce fait, le modèle d'*évaluation* permet d'explorer des solutions de diagnostic de décisions erronées qui n'ont pas été forcément envisagé par les enquêteurs.

Enfin d'un point de vue informatique, le filtrage des solutions par les biais semble à première vue une solution efficace pour ne considérer que les scénarios pertinents sur les grand nombres de scénario générés. Nous verrons toutefois sur le second cas d'étude que ce n'est pas forcément le cas pour toutes les situations d'accident.

VII.4.2 Accident du vol 5148

Nous allons dans cette section discuter des résultats pour un filtrage sur l'accident du vol 5148 du mont St-Odile (voir section VI.1) avec un seuil à $\mathcal{K} = 1$. Tout d'abord sans filtrage (*i.e* $\mathcal{K} = 0$) , nous trouvons 288 scénarios. Avec $\mathcal{K} = 1$, nous trouvons 240 scénarios, c'est-à-dire qu'une majorité des scénarios sont gardés. Nous allons discuter pour chaque pas de temps pourquoi nous trouvons un tel résultat. Tout comme l'accident précédent, nous omettons dans les exemples d'ignorances les propositions d'*inertie* qui sont ignorées par *extrapolation* ou par *mise à jour*.

VII.4.2.a Biais du premier pas de temps

Au premier pas de temps, les ignorances peuvent être expliquées par le biais d'optimisme, de faux souvenirs, d'évitement et de facilitation. Prenons deux exemples d'ignorances qui illustrent ces explications :

$$\begin{aligned} \mathcal{I}_1^a &= \{\mathrm{goodX}_1, \mathrm{goodY}_1, R_1^h\} \\ \mathcal{I}_1^b &= \{\mathcal{D}(\mathrm{goodX}_1), \mathcal{D}(\mathrm{goodY}_1), \mathrm{onVS}_1, \mathrm{onVS}_0\} \end{aligned}$$

L'agent fait face à deux observations qui sont chacune en contradictions avec un désir de l'agent. En ignorant les observations dans $\mathcal{I}_1^a$, l'agent fait un biais d'évitement car il n'observe pas les informations qui lui permettent de conclure que ses désirs ne sont pas satisfaits. Dans le cas de $\mathcal{I}_1^b$, ignorer les désirs est attendu du fait que l'agent observe des informations en contradiction avec ses

désirs, donc les deux ignorances des désirs sont considérées comme explicables. De plus, l'ignorance de la règle R_1^h dans $\mathcal{I}_1^a$ correspond à un biais d'optimisme du fait que l'agent, en l'ignorant évite que son désir goodY_2 ne soit pas satisfait. Enfin, dans $\mathcal{I}_1^b$, l'ignorance de onVS_1 peut être expliquée par un biais de facilitation de décision : l'agent porte son attention sur les désirs non satisfaits goodY_1 et goodX_1 et n'observe pas onVS_1. L'ignorance de onVS_0 est alors un faux souvenir car l'agent concentre son attention sur les désirs non-satisfaits.

Ainsi, les ignorances de l'agent correspondront finalement toujours à éviter les désirs non satisfaits ou à les prendre en compte et par conséquent à tomber dans le biais d'évitement ou de mettre à jour les désirs. Les seules ignorances qui sont filtrées sont du type :

$$\mathcal{I}_1^c = \{\mathrm{goodX}_1, \mathrm{goodY}_1, \mathrm{onVS}_1, \mathrm{onVS}_0\}$$

car dans $\mathcal{I}_1^c$ l'ignorance de onVS_0 ne peut pas être un faux souvenir car l'agent ne peut pas avoir son attention sur une autre observation qui ne satisfait pas un désir du fait que les observations sont ignorées.

En conclusion, nous trouvons bien l'explication du BEA pour l'accident du mont St-odile à savoir les ignorances $\mathcal{I}_1^b$: l'agent porte son attention sur la position horizontale de l'appareil et oublie qu'il est en mode *Vertical Speed*. Toutefois, du fait que la majorité des ignorances correspond à éviter un désir, une majorité d'ignorances n'est pas filtrée.

VII.4.2.b Biais du deuxième pas de temps

Pour le deuxième pas de temps, nous tombons finalement sur une situation similaire qu'au premier pas de temps. L'agent a le choix de considérer que ses désirs ne sont pas satisfaits au vu des observations ou ne pas observer les informations. Là encore, cela revient à un biais d'évitement ou une mise à jour des désirs. Tout comme le premier pas de temps, l'agent peut porter son attention sur les désirs qui ne sont pas satisfaits et ainsi ignorer l'observation du mode d'affichage :

$$\mathcal{I}_2^d = \{\mathcal{D}(\mathrm{goodX}_1), \mathcal{D}(\mathrm{goodY}_1), \mathrm{onVS}_2\}$$

Tout comme le premier pas de temps, les ignorances peuvent être expliquées par le fait que l'agent évite que ses désirs ne soit pas satisfaits (un biais d'évitement ou un biais d'optimisme). Ainsi, au deuxième pas de temps, tous les ensembles d'ignorances sont gardés.

VII.4.2.c Conclusion sur le vol 5148

Nous trouvons le scénario qui est considéré comme le plus plausible par le BEA avec les mêmes explications : le pilote oublie le mode de l'autopilote du fait qu'il se concentre sur la position horizontale de l'appareil. Toutefois, le filtrage dans cette

situation est bien moins efficace que sur le cas de l'accident du vol Rio-Paris. Ici seul environ 17% des scénarios sont filtrés contre environ 99.6% dans le cas du vol 447. Nous trouvons deux explications possibles pour comprendre ce filtrage moins efficace :

— La définition du biais d'optimisme et d'évitement est trop générale et englobe beaucoup trop de situations.
— La plausibilité d'un scénario ne peut pas être évalué que par la présence ou non des biais cognitifs. Par exemple si un état est expliqué par un biais cognitifs mais ne permet d'expliquer l'intention de l'action effectuée alors l'état ne doit pas être considéré comme plausible.

Nous allons discuter plus précisément de cette dernière possibilité dans la section suivante.

VII.5 Conclusion et limites

Nous avons pu dans ce chapitre proposer un modèle dit d'*évaluation* qui permet de déterminer une plausibilité sur les états de croyances retournés par le modèle dit d'*explication*. Pour ce faire, nous avons défini une taxonomie formelle des biais basée sur des caractéristiques définies grâces aux ignorances, états de croyances et choix de l'agent. À partir de ces caractéristiques, nous définissons les biais comme une conjonction de caractéristiques que respecte une ignorance φ. Un état de croyances plausible est alors un état de croyances dont une part des ignorances au-dessus d'un certain seuil $\mathcal{K}$ correspondent à des biais.

Nous avons pu montrer que l'utilisation des biais comme des filtres sur les états de croyances est utile d'un point de vue computationnel du fait qu'un nombre important de scénarios est écarté, mais aussi d'un point de vue de la compréhension de l'accident, car les scénarios conformes aux enquêtes sont gardés.

Bien que nous ayons montré que filtrer par les biais est efficace, il reste encore trop de scénarios plausibles pour qu'un utilisateur humain puisse les analyser. Il est donc nécessaire de filtrer encore plus les scénarios pour que notre modèle soit pertinent pour de l'aide à l'analyse d'accident. Nous pensons qu'il existe deux limitations dans notre modèle, toutes les deux en rapport avec la plausibilité, qui permettraient de rendre le filtrage plus efficace.

La première de ces limitations est mise en avant par le cas d'étude de l'accident du mont Saint-odile : évaluer la plausibilité d'un scénario uniquement par le prisme des biais détectés ou non n'est pas suffisant dans certaines situations. Nous pensons que certaines analyses supplémentaires doivent être effectuées pour évaluer plus précisément la plausibilité d'un scénario. Tout d'abord, nous ne prenons pas en compte les enchainement des biais dans un scénario qui pourraient être un indice de plausibilité. Par exemple, un biais d'attention sur un non-désir puis un biais d'évitement sur ce même désir ne semble pas représenter un comportement plausible. En effet, cela voudrait dire que l'agent porte toute son attention

sur un danger puis décide de ne plus du tout y porter attention. Il y a donc un manque de continuité dans le comportement de l'agent, ce qui est moins plausible qu'un scénario où l'agent continue de garder son attention sur le danger. Il est donc nécessaire de prendre en compte des explications de plus haut-niveaux en étudiant l'enchaînement des biais trouvés dans le temps. De plus, nous pensons qu'une analyse supplémentaire est à effectuer au niveau de l'intention de l'agent. En effet, par exemple, dans le cas du mont St-odile, les scénarios où l'agent tombe dans un biais d'évitement en n'observant pas qu'il est en mauvaise position verticale ne sont pas plausibles du fait qu'il n'y a aucune intention derrière l'action de rentrer la valeur 33 dans l'autopilote. En effet, l'agent n'ayant pas son désir d'être en bonne position verticale non-satisfait, il n'y a pas l'intention de corriger sa position derrière l'action effectuée. Il nous semble donc nécessaire d'analyser si l'état de croyances de l'agent permet d'expliquer aussi l'intention de l'agent afin de le considérer plausible ou non.

La deuxième limitation réside, comme nous avons pu le voir, dans le fait que certaines ignorances pouvaient être expliquées par plusieurs biais cognitifs. Par exemple, une ignorance peut être à la fois un biais de confirmation et un biais d'attention si l'agent ne prend pas en compte une information qui ne va pas dans le sens de ces anciennes croyances et qui en plus est contradictoire avec une information qui indique un danger. Le fait qu'une ignorance a plusieurs explications possibles ne peut que renforcer sa plausibilité. En effet, cela indique qu'il existe plusieurs stratégies cognitives qui ont pu mener à cette ignorance. Par conséquent, l'agent a plus de chance d'être tombé dans ce scénario, car il a plus de chance d'avoir utilisé une stratégie qui résulte de ces ignorances. Il est donc nécessaire que la plausibilité augmente pour un état de croyances dans un tel cas.

Toutefois, nous avons pu montré que le modèle d'*évaluation*, à travers son application sur différents accidents, est une base solide pour déterminer des diagnostics de décisions humaine erronées et explorer des diagnostics non envisagé par les experts. Les limitations abordées dans cette section sont des points d'améliorations de ce modèle d'*évaluation* et nous allons aborder des premières idées pour y répondre dans la prochaine section.

VIII - Conclusion et perspectives

VIII.1 Conclusion

Le travail effectué au cours de cette thèse a permis de poser les bases d'une approche formelle pour le diagnostic d'erreurs de prise de décision d'un humain.

L'étude de l'existant dans le domaine des sciences humaines, que nous avons présenté chapitre II, a montré que la problématique du diagnostic de l'erreur humaine peut être découpée en deux étapes. La première consiste à se placer du point de vue de l'agent humain dans une situation donnée et de déterminer les croyances de l'agent qui sont cohérentes avec sa décision. La deuxième étape consiste à trouver une explication à ces croyances. Pour cela, la littérature en sciences humaines s'appuie sur des explications à base de biais cognitifs qui sont des raisonnements erronés qui résultent d'heuristiques que les humains utilisent. Au cours de cette thèse, nous avons proposé une formalisation pour chacune de ces deux étapes.

La formalisation de ces deux étapes nous a conduits à résoudre plusieurs difficultés. La première consiste à représenter formellement les croyances et actions d'un agent dans un environnement dynamique. En effet, avant de diagnostiquer les prises de décisions erronées de l'agent, il est essentiel de représenter la situation dans laquelle se trouve l'agent au moment d'une décision. Nous avons pour cela étudié les différentes logiques de croyances et d'actions ainsi que les différentes opérations de changements de croyances permettant de représenter un agent rationnel dans un environnement dynamique. Cette étude de l'existant a été présentée section III.1. La deuxième difficulté consiste à diagnostiquer un système basé sur une logique de croyances et d'actions afin de déterminer les croyances de l'agent cohérentes avec sa décision. Nous nous sommes intéressés pour cela aux différentes approches formelles de diagnostic possible présentes dans la littérature que nous avons présenté section III.2. Enfin, pour déterminer les explications possibles à ces croyances, nous avons proposé une formalisation des biais cognitifs dans un langage logique similaire à celui utilisé pour la représentation des croyances. Nous avons pour cela étudié les différents travaux dans la littérature informatique qui

s'intéresse à capturer les biais cognitifs décisionnels dans un cadre formel et nous les avons présentés section III.3.

Ce travail bibliographique nous a permis de nous positionner sur deux domaines de connaissance : les sciences humaines et l'informatique. Nous avons choisi d'adopter la vision de [Dekker, 2006] pour la recherche des croyances de l'agent. Pour l'explication des croyances grâce au biais cognitifs, nous avons fait le choix de ne pas nous appuyer sur une taxonomie existante du fait du manque de consensus et des caractéristiques imprécises des différentes taxonomies qui rend difficile le travail de formalisation. Les choix que nous avons fait pour notre modèle informatique ont été largement orientés par la vision de Dekker et la recherche de cohérence entre les actions et les croyances. Nous avons fait notamment le choix de nous inspirer d'une logique de type BDI pour deux raisons. Le choix de l'utilisation d'une logique qui s'inspire d'une logique BDI de type « base de données » [Shoham, 2009] permet d'avoir des concepts proches de la cognition humaine et a une approche similaire à la vision de Dekker : la cohérence entre la base d'intention (*i.e* les actions) et la base de croyances. Dans la même optique, l'approche de *consistency-based diagnosis* [Reiter, 1987] a été choisie comme approche de diagnostic car elle permet de déterminer un diagnostic d'un système en retrouvant la cohérence entre la description du système et les observations faites sur ce système. Ainsi, nous retrouvons là encore l'esprit de la vision de Dekker. Enfin, pour représenter les biais, nous avons fait le choix de nous inspirer du travail de [Dutilh Novaes et al., 2016] qui repose sur un opérateur de révision de croyance pour capturer le biais de croyance. La révision de croyance pouvant être rapprochée à un *consistency-based diagnosis* comme vu section III.2, nous pensons que ces travaux sont un bon point de départ pour capturer les biais tout en restant dans la logique de la recherche de cohérence des croyances. Ces différents choix sont illustrés à travers les modèles dit d'*explication* et d'*évaluation* développés dans cette thèse.

VIII.1.1 Première contribution : le modèle d'explication

Le modèle d'*explication* que nous avons proposé consiste à déterminer les croyances possibles de l'agent qui permettent d'expliquer une prise de décision erronée. Celui-ci repose sur le principe de la recherche des ignorances possibles de l'agent pour retrouver la cohérence entre ses croyances et sa décision, sur le même principe que le *consistency-based diagnosis*. Pour cela, l'agent peut ignorer des observations, des désirs, des règles de raisonnement, des croyances précédentes ou l'inertie des croyances pour retrouver la cohérence.

L'originalité de ce modèle réside tout d'abord sur l'application d'une approche *consistency-based diagnosis* pour diagnostiquer une décision humaine. Bien que l'approche soit classique pour diagnostiquer un système, à notre connaissance, celle-ci n'a jamais été appliqué sur cette problématique. Cela nous a amené à

soulever de nouvelles problématiques relatives au diagnostic de l'erreur humaine comme le fait de prendre en compte des erreurs sur les différents problèmes d'incohérences auquel l'agent doit faire face. Ces erreurs peuvent être liées à une révision de croyances, une action incohérente ou des erreurs liées à l'*inertie* des croyances (*i.e* des distorsions).

L'une des originalités de notre modèle et de notre algorithme de diagnostic est donc de prendre en compte cette variété d'erreurs. En effet, les ignorances possibles sont construites en résolvant, successivement par des *Minimal Correction Sets*, les différents problèmes d'incohérences auquel l'agent doit faire face (révision, action incohérente, extrapolation, mise à jour et distorsion). Ce diagnostic par itération nous permet de capturer la variété des erreurs possibles chez un agent humain. De plus, cela permet une expressivité sur l'explication plus importante du fait qu'une ignorance peut être rattachée à une nature différente en fonction du problème d'incohérence qu'elle résout. Par exemple, ignorer une croyance pour un problème de révision de croyance traduit une préférence envers une autre croyance. Nous pouvons alors exprimer des diagnostics plus précis pour un analyste humain. Par exemple : *la révision de telle croyance a conduit à telle extrapolation qui a eu pour conséquence le choix d'action*.

L'implémentation du modèle d'*explication* nous a permis de tester le modèle sur l'accident du vol 447 de Rio-Paris et du vol 5148 du Mont Sainte-odile. Cette application du modèle sur ces deux accidents nous a permis de confirmer que certains scénarios trouvés par le modèle d'explication était cohérent avec les rapports des enquêteurs des accidents.

VIII.1.2 Deuxième contribution : validation théorique et implémentation de notre opérateur de diagnostic

Nous avons pu montrer section VI.4 que l'ensemble des problèmes d'incohérences auquel l'agent fait face peut être résolu par un opérateur de révision de croyance respectant les axiomes d'AGM. Afin de valider que notre opérateur de diagnostic peut être utilisé pour résoudre l'ensemble de ces problèmes, nous avons pu utiliser l'assistant de preuve Isabelle afin de prouver formellement que calculer les *Minimal Correction Sets* permet de construire des opérateurs de révision de croyance respectant AGM.

Cette validation de la pertinence de notre opérateur de diagnostic pour résoudre l'ensemble des problèmes d'incohérences de notre problématique, l'utilisation des MCSes permet de profiter de l'ensemble des algorithmes de la littérature pour calculer ces ensembles. Nous avons proposé dans cette thèse une implémentation de l'algorithme de [Liffiton et al., 2008] à l'aide d'un SMT-solver pour le calcul des MCSes. L'opérateur de diagnostic peut donc profiter des avancées sur les algorithmes de calcul des MCSes ainsi que des avancées sur les SMT-solver pour

améliorer les performances de cet opérateur. Cette équivalence entre notre opérateur de diagnostic et un opérateur AGM permet donc de faire le lien entre la théorie et une implémentation concrète d'un opérateur AGM.

VIII.1.3 Troisième contribution : correction et complétude de l'algorithme de diagnostic

Afin de valider que l'algorithme de diagnostic utilisé dans le modèle d'*explication* retourne des diagnostics aussi pertinents qu'un *consistency-based diagnosis*, nous avons utilisé l'assistant de preuve Isabelle pour montrer la correction et la complétude de l'algorithme. Nous avons montré que l'algorithme de diagnostic trouve des solutions similaires et ne passe pas à côté d'une solution trouvée par un *consistency-based diagnosis*.

Au-delà de cette validation, la preuve de correction et complétude a permis de montrer que certaines solutions retrouvées par l'algorithme ne sont pas minimales. Par conséquent, les états de croyances résultant de l'algorithme peuvent avoir ignoré plus de choses que nécessaire, ce qui augmente la complexité des comportements trouvés. Par exemple, un agent peut effectuer une révision de croyance incompatible avec sa décision. Cette révision conduit à un ensemble de corrections qui n'est pas minimal globalement mais elle traduit aussi des mécanismes d'erreurs humaines spécifiques. Ainsi, notre algorithme permet de capturer une plus large palette d'erreurs que ne le ferait un *consistency-based diagnosis* tout en s'assurant que chaque solution trouvée pour chaque problème d'incohérence est minimale.

VIII.1.4 Quatrième contribution : Une première proposition de taxonomie formelle des biais

Afin de formaliser les biais cognitifs et d'expliquer les états de croyances trouvés par notre modèle, nous avons développé une taxonomie formelle permettant de caractériser les biais. Cette taxonomie se base sur les ignorances, les états de croyances, les choix possibles de l'agent ainsi que trois grains d'analyse. Chaque grain d'analyse permet de mettre en perspective une ignorance par rapport à la nature et au type de l'ignorance, aux autres choix possibles ou aux états de croyances précédents. Les caractéristiques de chaque grain d'analyse peuvent être trouvées section VII.1. Nous avons alors proposé de définir chaque biais cognitif par un ensemble de caractéristiques et nous avons illustré ce principe sur huit biais principaux. Ces définitions permettent de valider le potentiel explicatif du cadre formel proposé par notre modèle d'explication, capable d'exprimer un éventail de biais de natures différentes : les biais d'attention, de confiance et de mémoire.

Au-delà de l'utilité de cette taxonomie pour définir les biais cognitifs, dans notre modèle informatique, celle-ci peut être exploitée dans le domaine des sciences humaines pour répondre aux limites des différentes taxonomies des biais dans la

littérature. C'est-à-dire des caractéristiques imprécises sujettes à interprétation qui par conséquent entraînent une distinction des biais difficile. Le langage formel ne permettant pas différentes interprétations possibles, la distinction entre les biais dans une taxonomie formelle ne peut être que définie ou ne pas exister. Nous pensons donc que la taxonomie formelle proposée dans cette thèse est un premier pas allant dans ce sens.

VIII.1.5 Cinquième contribution : le modèle d'évaluation

Le modèle d'*évaluation* consiste, à partir de l'identification de biais cognitifs dans les ignorances, à déterminer la plausibilité des états de croyances et de filtrer les scénarios pour ne garder que les scénarios les plus plausibles. Pour cela, le modèle d'*évaluation* prend en entrée l'ensemble des scénarios trouvés par le modèle d'*explication* et un ensemble de biais définis par les caractéristiques de la taxonomie formelle. Une ignorance est alors considérée comme « explicable » si elle correspond à la définition d'un biais cognitif ou d'une ignorance attendue chez un agent rationnel (*i.e* extrapolation ou mise à jour).

L'originalité de ce modèle est de se servir des biais cognitifs comme une heuristique pour déterminer les scénarios les plus plausibles. Les travaux de la littérature en informatique utilisent les biais cognitifs comme une aide à la prédiction d'une décision erronée. Dans nos travaux, nous les utilisons comme une aide à la compréhension d'une décision erronée : plus un scénario peut être expliqué par des biais, plus ce scénario est plausible.

Le modèle prend en paramètre un seuil $\mathcal{K}$ pour ne considérer que les scénarios qui sont plausibles au-dessus de ce seuil, c'est-à-dire que la proportion d'ignorances explicables par un biais est supérieure à K. Ainsi un état de croyance est considéré comme plausible si au moins $\mathcal{K}\%$ des ignorances de l'état sont explicables. Un scénario est considéré comme plausible si chaque état de croyance du scénario est plausible avec le seuil $\mathcal{K}$. Le modèle d'*évaluation* permet donc à l'utilisateur d'explorer ou non selon son choix des scénarios considérés comme plus ou moins plausibles. L'utilisation de ce seuil a plusieurs avantages. Le premier est de limiter l'affichage des scénarios pour ne présenter que les plus plausibles à l'utilisateur et ainsi faciliter son travail d'investigation. Le deuxième est de pouvoir laisser le choix à l'utilisateur d'explorer des scénarios dont aucune explication n'est trouvée avec les définitions des biais en paramètre du modèle. Une telle exploration permet notamment de définir des nouveaux biais empiriquement en repérant des ignorances non explicables qui sont cohérentes avec un biais dans la littérature mais qui n'est pas encore défini dans le modèle. Ainsi, nous avons proposé, en nous appuyant sur les théories en SHS, un modèle informatique pour l'analyse de l'erreur humaine, nous l'avons implémenté et nous avons intégré des mécanismes pour manipuler les résultats du diagnostic. Ce modèle offre de nombreuses perspectives que nous

présentons dans la section suivante.

VIII.2 Perspectives

Nous avons choisi de regrouper les perspectives de nos travaux en trois groupes : celles qui relevent d'améliorations immédiates de notre modèle, celles qui nécessitent un travail de recherche plus approfondi et enfin les nouvelles questions de recherche soulevées par nos résultats.

VIII.2.1 Perspectives à court-terme

Nous pensons que plusieurs objectifs sont réalisables rapidement pour augmenter l'expressivité du modèle et par conséquent les biais détectés.

VIII.2.1.a Ajouts de nouveaux biais

Une première perspective à court-terme est de définir d'autres biais à partir de la taxonomie. Nous avons dans cette thèse défini les biais qui nous semblaient les plus essentiels dans l'analyse d'une décision erronée dans un accident d'aviation. Les huit biais que nous avons définis ne sont pas exhaustifs, la littérature scientifique en dénombrant 151 comme nous avons vu section II.3. Il serait donc intéressant d'agrémenter le modèle d'*évaluation* d'autres biais afin d'augmenter les explications possibles pour les scénarios et offrir des scénarios plus variés. Pour cela, deux méthodologies sont possibles. La première consiste à reprendre l'approche suivie dans cette thèse, c'est-à-dire se concentrer sur les biais cognitifs qui ont été identifiés dans la littérature dans les accidents et de les définir à partir des caractéristiques de la taxonomie. Cette méthode a l'avantage de s'appuyer sur des exemples concrets pour valider les définitions des nouveaux biais, tout en disposant de descriptions détaillées des situations dans lesquelles les biais ont été identifiés. Par exemple, la collision de Tenerife en 1977 [Board, 1979] et l'accident nucléaire du Three Mile Island en 1979 [Commission, 1979] sont deux bons points de départ. La deuxième méthode, plus exploratoire, consiste à modéliser des accidents ou des situations de prises de décision erronées et d'explorer les scénarios qui contiennent des ignorances qui ne sont pas expliquées pour les rattacher potentiellement à des biais dans la littérature. Toutefois, cette méthode nécessite une vision experte des biais du fait que cette méthode nécessite au préalable une connaissance sur l'ensemble des biais possibles.

VIII.2.1.b Ajouts des émotions

Notre modèle d'*évaluation* dans sa version actuelle, les émotions qui sont un élément important dans les comportements humains, comme l'a montré la

nombreuse littérature sur le sujet en SHS [Lazarus et al., 1984, Frijda, 1988, Lerner et al., 2015]. Le concept qui se rapproche des émotions dans notre modèle sont les caractéristiques liées à un désir insatisfait ou satisfait que nous considérons comme une perte ou un gain pour l'agent. Nous pouvons les voir comme des émotions négatives ou positives. Or le spectre des émotions est bien plus grand qu'un stimulus négatif ou positif avec par exemple la peur, l'espoir, la joie, le regret, la déception, etc. Toutes ces émotions jouent un rôle important dans la prise décision et ne peuvent être mises de côté pour expliquer une décision. Il nous semble donc nécessaire de les intégrer afin d'évaluer la possibilité d'un état de croyances et biais de l'agent combiné avec l'état émotionnel de l'agent à un instant t. Par exemple, il a été montré qu'il y a une propension au biais de facilitation d'attention lors d'un stimulus de danger [Cisler et al., 2010], il est donc attendu que l'agent ayant un biais de facilitation éprouve de la peur. Dans ce cas, nous considérons les scénarios où l'agent a un biais de facilitation et ressent de la peur comme plus probable que le scénario où l'agent a un biais de facilitation sans ressentir de la peur.

Un bon point de départ pour prendre en compte ces émotions dans notre modèle est d'utiliser les travaux de formalisation des émotions de [Adam et al., 2009]. La formalisation d'Adam se base sur le modèle des émotions OCC de [Ortony et al., 1990] et sur une logique de type BDI afin de modéliser 20 émotions. Par exemple, l'émotion de la joie pour l'agent i est exprimé par :

$$\mathrm{Joy}_i\,\varphi \quad \stackrel{\text{def}}{=} \quad \mathrm{Bel}_i\,\varphi \wedge \mathrm{Des}_i\,\varphi$$

c'est-à-dire que l'agent i désire φ et croît φ. Du fait que la logique repose sur des concepts BDI proche de la logique utilisée dans notre approche, le modèle proposé par Adam, Herzig et Longin est une logique de choix pour exprimer les émotions dans notre modèle. Toutefois, l'intégration de cette logique dans notre modèle nécessite une adaptation. En effet, il existe un opérateur $\mathrm{Prob}_i\,\varphi$ dans leur modèle qui permet de construire des émotions et qui exprime le fait que : L'agent i croit que ϕ est plus probable que $\neg\phi$ (un opérateur moins fort que la croyance : $\mathrm{Bel}_i\,\varphi \rightarrow \mathrm{Prob}_i\,\varphi$ mais pas l'inverse). Par exemple, la peur est exprimée par :

$$\mathrm{Fear}_i\,\varphi \quad \stackrel{\text{def}}{=} \quad \mathrm{Prob}_i\,\varphi \wedge \neg\,\mathrm{Bel}_i\,\varphi \wedge \mathrm{Des}_i\,\neg\varphi$$

L'agent i ne croit pas en φ mais pense que φ est probable et ne le désire pas.

Il pourrait donc être intéressant d'étendre notre modèle pour y intégrer cette notion de « croyance probable » afin de capture des émotions.

VIII.2.1.c Prise en compte des raisonnements probabilistes

Le langage logique utilisé dans nos travaux ne permet pas d'exprimer des raisonnements probabilistes. Par exemple, nous ne pouvons pas définir une règle qui dit que s'il y a des nuages, alors il est probable à 60% qu'il y ait de la pluie. Certains biais cognitifs nécessitent un tel cadre probabiliste, en particulier dans le

contexte de la prise de décision sous incertitude, comme l'*aversion au risque* qui est la tendance des individus à prendre une décision ayant pour effet un gain peu important mais très probable pour l'agent plutôt qu'une décision ayant pour effet un gain très important mais peu probable [Kahneman et al., 1979]. C'est le cas aussi pour tous les biais statistiques que l'on retrouve beaucoup dans la prise de décision médicale [O'Sullivan et al., 2018]. Par exemple, l'oubli de la fréquence de base peut pousser un médecin après un test à déclarer un jeune patient atteint d'une maladie alors que la fréquence de base de cette maladie est très basse sur la population jeune. Le test est donc sûrement un faux positif. Pouvoir modéliser les raisonnements probabilistes permet de prendre en compte à la fois les émotions, d'étendre le champ des biais possibles pouvant être exprimés et dans le même temps appliquer l'approche à d'autres domaines que l'aviation (*e.g* la médecine, l'économie, etc).

La littérature en logique s'est largement intéressée à l'intégration des probabilités dans un langage logique [Hójek, 2017, Demey et al., 2013]. Nous pouvons donc nous inspirer de tous ces travaux pour intégrer les raisonnements probabilistes dans notre approche. Nous pensons toutefois qu'une première solution à explorer est une logique probabiliste qualitative telle que proposée dans la formalisation de [Adam et al., 2009] où un opérateur exprime le fait qu'une proposition est plus probable que sa négation sans donner une quantité précise sur cette probabilité. Utiliser un tel opérateur permet une adaptation plus facile de la logique qualitative utilisée dans nos travaux pour intégrer les probabilités. En ce sens, nous avons commencé à explorer une première solution s'inspirant de cet opérateur.

Cette solution consiste à différencier deux types de règles de raisonnements :

— Des règles dites « fortes » qui sont équivalentes aux règles de raisonnement de notre approche : toute croyance déduite par une règle forte est considérée comme une croyance au sens d'Adam. C'est-à-dire que si φ est déduit uniquement par des règles fortes, alors c'est équivalent à $\mathrm{Bel}_i\,\varphi$.

— Des règles dites « faibles » qui permettent de déduire qu'une croyance est probable. Par exemple, si $\mathrm{cloud}_t \rightarrow \mathrm{rain}_{t+1}$ est dans les règles faibles, cela exprime le fait que s'il y a des nuages, l'agent pense qu'il est probable qu'il y ait de la pluie au temps suivant. Ainsi si une croyance φ est déduite par au moins une règle faible, alors c'est équivalent à $\mathrm{Prob}_i\,\varphi$.

Toutefois, cette approche encore préliminaire n'a pas été présentée dans ce manuscrit et fera l'objet de publications futures.

VIII.2.2 Perspectives à moyen-terme

Nous pensons qu'à moyen-terme, le principal objectif à atteindre est d'améliorer l'algorithme d'évaluation afin que les scénarios présentés à l'utilisateur du modèle soient les plus pertinents.

VIII.2.2.a Étendre l'évaluation des scénarios

Dans cette thèse, nous avons évalué la plausibilité d'un scénario en fonction de la proportion des ignorances qui pouvait être expliquée par un biais cognitif. Nous avons pu voir que bien que filtrés les scénarios les plus plausibles avec cette méthode pouvaient être efficaces mais parfois insuffisants (voir section VII.4). Nous pensons qu'il existe plusieurs moyens d'étendre et d'améliorer l'évaluation des scénarios. Une première piste est d'effectuer une « méta évaluation » des biais identifiés dans un scénario. C'est-à-dire évaluer l'agencement des biais dans les scénarios. Cette évaluation peut être faite sur un pas de temps. Par exemple, si l'agent ignore une observation φ qui correspond à un biais de facilitation sur le désir d et à la fois ignore une autre observation ψ qui correspond à un biais d'évitement sur ce même désir d, alors il y a un problème de cohérence dans la stratégie utilisée du fait que l'agent porte et ne porte pas à la fois son attention sur d. Un tel état de croyance devrait être donc considéré comme moins plausible.

L'évaluation de l'agencement des biais peut être effectuée aussi à travers les pas de temps du scénario. Par exemple, considérons un scénario où un agent décide de faire preuve d'optimisme en ignorant une règle de raisonnement qui lui dit qu'une alarme lui indique un danger. Il décide ensuite d'ignorer l'alarme dans les pas de temps suivant, ce qui correspond à un biais d'évitement, pour à nouveau ignorer l'alarme au troisième pas de temps (biais d'optimisme). Dans un tel scénario, l'agent change de stratégie plusieurs fois sans que ce soit nécessaire pour maintenir la cohérence. Si l'agent utilise un biais d'optimisme au premier pas de temps, il semble plus plausible qu'il applique le même biais au pas de temps suivant plutôt qu'entremêler des biais différents, du moins si l'on s'en tient au principe philosophique du rasoir d'Ockham : l'explication la plus simple est souvent préférable. Dans la même idée, le fait qu'un agent applique un biais de facilitation sur une information puis un biais d'évitement sur la même information au pas de temps suivant semble peu cohérent, d'autant plus que ces deux biais de raisonnement suivent des stratégies opposées. Nous pensons que l'évaluation de la plausibilité d'un scénario doit pouvoir évaluer ces agencements de biais dans le temps, notamment en évaluant négativement les agencements qui ne sont pas cohérents ou les agencements trop complexes alors qu'il existe une solution plus simple.

L'évaluation de l'agencement des biais doit pouvoir aussi considérer des états de croyances comme plus plausibles en fonction du nombre d'explications possibles pour une ignorance. Par exemple, il est possible qu'une ignorance soit à la fois un biais de facilitation et de confirmation : l'agent porte son attention sur une observation qui ne va pas dans le sens de ces désirs et en plus vient confirmer ce que l'agent croyait précédemment. Un tel état de croyance devrait avoir sa plausibilité augmentée fortement du fait que plusieurs biais cognitifs sont possibles pour expliquer une même ignorance : la probabilité que l'agent ait utilisé une des heuristiques qui résulte d'un de ces biais est plus grande. En d'autres termes, plus de biais sont possibles sur une ignorance, plus une erreur est possible de la part de

l'agent.

Pour étendre l'évaluation, il nous semble donc intéressant d'effectuer une « méta évaluation » des biais cognitifs identifiés est de déterminer leur cohérence dans un état de croyances et à travers le temps ainsi que leurs nombres pour expliquer une même ignorance.

Enfin, nous pensons qu'un ingrédient dans l'analyse des scénarios est manquant pour améliorer grandement l'évaluation des scénarios : l'intention de l'action. Pour effectuer une action, un agent doit avoir l'intention de le faire, c'est-à-dire qu'il existe une raison qui pousse l'agent à cette action. Par exemple, dans le cas des logiques BDI, l'action est effectuée pour satisfaire un désir. Dans notre approche, nous n'avons pas besoin de l'intention de l'action dans le modèle d'*explication* du fait que nous savons que l'action a été effectuée et nous cherchons seulement les croyances cohérentes avec cette action. Toutefois, nous pensons que le modèle d'*évaluation* profiterait de la prise en compte de l'intention de l'action. En effet, dans le cas de l'accident du Mont Sainte-Odile, notre modèle produit un scénario où l'absence d'intention est problématique : celui où l'agent ignore l'observation sur sa position verticale. En effet, cette action a pour effet de corriger la position verticale. Or si l'agent a ignoré l'information, il n'a aucune raison d'effectuer une correction ! Nous pensons donc que l'état de croyances qui résulte de cette ignorance ne peut pas être considéré comme plausible du fait que l'intention de l'action effectuée est manquante. Nous pensons donc qu'il serait intéressant de prendre en compte l'intention de l'action dans l'évaluation des états de croyances. Un bon point de départ serait de se baser sur les différentes représentations de l'intention dans les logiques BDI que nous avons abordé sous-section III.1.4.

VIII.2.2.b Validation des scénarios par des experts

Nous avons pu voir section V.5 et section VII.4 que notre approche permet de retrouver des scénarios avec les mêmes explications que celles proposées par les enquêteurs de l'accident. Toutefois, d'autres scénarios pouvant être expliqués par des biais différents et considérés comme plausibles par le modèle d'*évaluation*. Se pose alors la question de la pertinence de ces autres scénarios. En d'autres termes, peut-on évaluer la validité des scénarios qui ne correspondent pas à la « meilleure » solution, c'est-à-dire le scénario retenu par les enquêteurs.

Pour cela, nous proposons de nous inspirer des méthodes d'entretien des sciences humaines, notamment de l'entretien *semi-directif* [Salah et al., 2018]. L'objectif serait de réaliser un entretien avec plusieurs experts (*e.g* des enquêteurs du BEA) où un accident leur serait présenté suivi des scénarios considérés comme plausibles par notre modèle d'*évaluation*. Les experts auraient pour tâches de classer les scénarios du plus plausible au moins plausible. De cette tâche suivra l'entretien *semi-directif* où les questions seront dirigées vers la justification du classement des différents scénarios.

Ces entretiens serviraient alors à déterminer si les scénarios plausibles retour-

nés par le modèle d'*évaluation* sont considérés comme pertinents et pourquoi. De plus, ces entretiens seraient une mine d'informations pour étendre l'évaluation des scénarios. En effet, les entretiens mettraient en avant les différents critères de comparaisons des hypothèses d'un scénario d'accident. Ces critères seraient potentiellement différents de ceux que nous avons mis en avant dans la section précédente ou seraient une confirmation de ces critères (*i.e* présence de plusieurs biais, rasoir d'Ockham, cohérence des biais, etc).

En conclusion, nous pensons que notre approche a tout à gagner de ces entretiens, que ce soit en termes de validation de l'approche mais aussi pour améliorer le modèle d'*évaluation*.

VIII.2.3 Perspectives à long-terme

Nos travaux ouvrent la voie à un vaste champ de problèmes pour l'analyse de l'erreur humaine à l'aide de modèles formels. Nous proposons ici deux pistes qui nous semblent prioritaires et particulièrement intéressantes à creuser : la prise en compte d'un contexte multi-agent dans le diagnostic et la conception d'un environnement virtuel d'analyse de l'erreur humaine.

VIII.2.3.a Passer à un contexte multi-agent

Notre approche s'intéresse à la prise de décision d'un unique agent dans un contexte mono-agent. Notre langage de modélisation ne permet pas de prendre en compte des situations où l'interaction sociale joue un rôle important dans l'explication de l'accident. Par exemple, dans l'accident de Tenerife [Board, 1979], c'est en partie l'interaction entre la tour de contrôle et les pilotes de l'avion qui explique l'accident. Considérer une situation d'accident multi-agent nécessiterait d'intégrer plusieurs concepts importants que nous allons illustrer à travers des exemples de biais cognitifs.

Un premier exemple est le phénomène de la paresse sociale c'est la tendance des individus à fournir moins d'effort proportionnellement à la taille du groupe [Karau et al., 1993]. Ainsi, une erreur de prise de décision du groupe peut ne pas être corrigée, car chaque individu du groupe pense qu'un autre individu la corrigera à sa place. Prendre en compte un tel biais nécessiterait de modéliser une *théorie de l'esprit* pour l'agent qui est modélisé. C'est-à-dire pouvoir modéliser les croyances de l'agent sur les désirs, croyances et intentions des autres agents. Par exemple, l'agent croit qu'un autre agent a l'intention de faire une action permettant de corriger l'erreur. Il existe plusieurs solutions dans la littérature permettant de modéliser la théorie de l'esprit en logique notamment grâce à la Dynamic Epistemic Logic [Van Ditmarsch et al., 2007, Dissing et al., 2020]. Ces logiques sont donc une première base solide pour prendre en compte la théorie de l'esprit dans notre approche. Toutefois, l'ajout de la théorie de l'esprit introduit de nouvelles questions sur les incohérences qui vont être introduites par la théorie de l'esprit. Par exemple,

l'agent va devoir potentiellement réviser ses croyances sur les croyances des autres agents ou avoir une action incohérente avec ce qu'il croit sur les intentions des autres agents. En plus des nouvelles incohérences à prendre en compte, la théorie de l'esprit introduit de nouveaux types d'ignorance qui doivent être considérés dans la taxonomie des biais. Ainsi la théorie de l'esprit permet d'exprimer des situations plus complexes, mais nécessite de gérer les nouvelles incohérences qui vont être introduites et exprimer les nouvelles caractéristiques dans la taxonomie afin d'exprimer des biais tel que la paresse sociale.

Un deuxième exemple est le biais de l'autorité qui est la tendance des individus à surévaluer l'opinion d'une personne ayant une autorité sur un sujet [Hinnosaar et al., 2012]. Par exemple, un copilote aura plus tendance à faire confiance à la décision du commandant de bord qui est son supérieur hiérarchique. Prendre en compte un tel biais nécessiterait de modéliser la confiance qu'un agent sur un autre agent (humain ou artificiel) : l'agent a a confiance en l'agent b et aura tendance à privilégier les informations provenant de b. Là encore, la littérature s'est déjà intéressée à la formalisation en logique de la confiance avec des logiques proches de la BDI [Herzig et al., 2010, Leturc et al., 2018] et offre une base solide pour intégrer la confiance dans notre approche. Cette intégration nécessitera aussi d'ajouter les caractéristiques propres aux ignorances liées à la confiance afin de pouvoir définir des biais liés à la confiance.

En conclusion, de nombreux travaux dans la littérature permettent d'adapter la logique utilisée dans notre approche pour l'agrémenter de concepts nécessaires à un contexte multi-agent. Toutefois, cela nécessitera un travail sur la gestion des nouvelles incohérences introduites par ces concepts et les nouvelles caractéristiques à définir dans la taxonomie.

VIII.2.3.b Exploration des scénarios

L'objectif principal de nos travaux est de pouvoir offrir un outil d'aide à la décision pour l'investigation d'un accident où un agent humain rentre dans la prise dé décision. Un utilisateur de notre modèle de diagnostic peut-être donc possiblement un enquêteur du BEA, c'est-à-dire des utilisateurs qui ne sont pas forcément sensibles aux méthodes formelles. Il y a donc un travail d'Interface Homme-Machine entre les scénarios en sortie du modèle et l'exploitation que l'utilisateur du modèle peut en faire. Les scénarios étant une suite d'état de croyances d'un agent, nous pensons qu'une piste à étudier pour explorer les scénarios est la réalité virtuelle. En effet, toujours dans la même approche que suivie dans cette thèse, la réalité virtuelle permet *une mise en situation* [Bowman et al., 2007, Slater, 2018], c'est-à-dire, prendre le point de vue de l'agent qui a effectué une prise de décision erronée. Par exemple, nous pourrions imaginer dans le cas de l'accident de Rio-Paris un environnement de réalité virtuelle où le cockpit de l'avion est représenté avec les outils de navigation au même état que pendant l'accident. À partir de là, plusieurs questions de recherches se posent pour la représentation des scénarios de notre

modèle dans cet environnement virtuel.

Une première question de recherche se pose sur la représentation des croyances dans un environnement de réalité virtuelle. Certaines croyances peuvent être rattachées à un phénomène physique observable, par exemple si l'agent croit en une alarme dans le cockpit, il suffit de représenter que l'alarme sonne dans l'environnement virtuel. Toutefois, certaines croyances ne peuvent pas être représentées physiquement. Par exemple, la croyance de la survitesse d'un avion est le fruit d'un raisonnement basé sur plusieurs informations dans le cockpit et est donc une construction de l'esprit d'un phénomène non directement observable. On retrouve la même chose pour les désirs : désirer ne pas être en survitesse n'a pas de représentation physique, mais est une construction de l'esprit. Il y a donc un besoin de représenter ce que l'agent croit de manière tangible et intelligible pour une mise en situation dans l'accident.

Une deuxième question de recherche se pose sur la représentation des ignorances de l'agent dans cet environnement virtuel. Les croyances de l'agent étant des ignorances sur le point de vue de l'agent (voir chapitre IV), représenter les croyances passe par la représentation des ignorances de l'agent. Le problème est que ces ignorances sont de natures différentes (*i.e* observations, règles, désirs, etc). Ainsi, représenter l'ignorance d'une observation pour un biais d'attention ne peut être équivalent à la représentation d'une ignorance d'une règle de raisonnement qui implique que certaines croyances ne seront pas déduites. Tout comme la représentation des croyances, il est nécessaire de rendre intelligible pour un utilisateur mis en situation dans l'accident ce que l'agent a ignoré pour effectuer une décision erronée.

Enfin, au-delà de la représentation des croyances et des ignorances à un instant t, l'interface proposée devra permettre d'explorer l'ensemble des scénarios plausibles retourné par le modèle d'*évaluation*. Chaque scénario étant une suite d'état de croyances, il est nécessaire de proposer une interface capable de comparer visuellement les différents états de croyances possibles à un instant t et de mettre en perspective les différents biais dans chaque état possible.

En conclusion, nous pensons qu'explorer la mise en situation des utilisateurs dans les scénarios retournés par notre approche permet à la fois un outil d'exploration plus facile à utiliser tout en abordant des questions de recherches d'IHM nouvelles et intéressantes.

Bibliographie

[Adam et al., 2009] Adam, C., Herzig, A., and Longin, D. (2009). A logical formalization of the occ theory of emotions. *Synthese*, 168(2) :201–248.

[Alchourrón et al., 1985] Alchourrón, C. E., Gärdenfors, P., and Makinson, D. (1985). On the logic of theory change : Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2) :510–530.

[Arnaud et al., 2017a] Arnaud, M., Adam, C., and Dugdale, J. (2017a). Les limites du BDI pour rendre compte du comportement humain en situation de crise. In *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA 2017)*.

[Arnaud et al., 2017b] Arnaud, M., Adam, C., and Dugdale, J. (2017b). The role of cognitive biases in reactions to bushfires. In *ISCRAM*, Albi, France.

[Arnott, 2001] Arnott, D. (2001). A taxonomy of decision biases.

[Aucher et al., 2013] Aucher, G. and Schwarzentruber, F. (2013). On the complexity of dynamic epistemic logic. *arXiv preprint arXiv :1310.6406*.

[Baltag et al., 1998] Baltag, A., Moss, L. S., and Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*, pages 43–56.

[Bastardi et al., 2011] Bastardi, A., Uhlmann, E. L., and Ross, L. (2011). Wishful thinking : Belief, desire, and the motivated evaluation of scientific evidence. *Psychological science*, 22(6) :731.

[BEA, 1993] BEA (1993). Bea f-ed920120. Technical report, Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile.

[BEA, 2012] BEA (2012). Bea f-cp090601. Technical report, Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile.

[Belardinelli et al., 2017] Belardinelli, F., Ditmarsch, H., and Hoek, W. (2017). A logic for global and local announcements. *Electronic Proceedings in Theoretical Computer Science*, 251 :28–42.

[Birnbaum et al., 2003] Birnbaum, E. and Lozinskii, E. L. (2003). Consistent subsets of inconsistent systems : structure and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence*, 15(1) :25–46.

[Board, 1979] Board, N. A. S. (1979). Final report and comments of the netherlands aviation safety board of the investigation into the accident with the collision of klm flight 4805, boeing 747-207b, ph-buf and pan american flight 1736, boeing 747-121, n736pa at tenerife airport, spain on 27 march 1977.

[Boutilier et al., 1995] Boutilier, C. and Beche, V. (1995). Abduction as belief revision. *Artificial intelligence*, 77(1) :43–94.

[Bowman et al., 2007] Bowman, D. A. and McMahan, R. P. (2007). Virtual reality : how much immersion is enough ? *Computer*, 40(7) :36–43.

[Bratman, 1987] Bratman, M. (1987). Intention, plans, and practical reason.

[Brusoni et al., 1998] Brusoni, V., Console, L., Terenziani, P., and Dupré, D. T. (1998). A spectrum of definitions for temporal model-based diagnosis. *Artificial Intelligence*, 102(1) :39–79.

[Buchanan et al., 1984] Buchanan, B. and Shortliffe, E. (1984). *Rule-based Expert System – The MYCIN Experiments of the Stanford Heuristic Programming Project*.

[Camerer et al., 2004] Camerer, C. and Loewenstein, G. (2004). Behavioral economics : Past, present, future. *Advances in Behavioral Economics*.

[Carter et al., 2007] Carter, C. R., Kaufmann, L., and Michel, A. (2007). Behavioral supply management : a taxonomy of judgment and decision-making biases. *International Journal of Physical Distribution & Logistics Management*.

[Ceschi et al., 2019] Ceschi, A., Costantini, A., Sartori, R., Weller, J., and Di Fabio, A. (2019). Dimensions of decision-making : An evidence-based classification of heuristics and biases. *Personality and Individual Differences*, 146 :188 – 200.

[Cisler et al., 2010] Cisler, J. M. and Koster, E. H. (2010). Mechanisms of attentional biases towards threat in anxiety disorders : An integrative review. *Clinical psychology review*, 30(2) :203–216.

[Cohen et al., 1990] Cohen, P. R. and Levesque, H. J. (1990). Intention is choice with commitment. *Artificial intelligence*, 42(2-3) :213–261.

[Commission, 1979] Commission (1979). Three mile island : a report to the commissioners and to the public. volume i.

[Commission, 1986] Commission, P. (1986). *Report to the President by the Presidential Commission on the Space Shuttle Challenger Accident*. Number vol. 1 in Report to the President by the Presidential Commission on the Space Shuttle Challenger Accident. The Commission.

[Conversy et al., 2014] Conversy, S. and al. (2014). L'accident du vol AF447 Rio-Paris, un cas d'étude pour la recherche en IHM. In *IHM'14, 26e conférence francophone sur l'Interaction Homme-Machine*, pages 60–69. ACM.

[Cooper et al., 2021] Cooper, M. C., Herzig, A., Maffre, F., Maris, F., Perrotin, E., and Régnier, P. (2021). A lightweight epistemic logic and its application to planning. *Artificial Intelligence*, 298 :103437.

[Dastani et al., 2012] Dastani, M. and Lorini, E. (2012). A logic of emotions : from appraisal to coping. In *AAMAS*, pages 1133–1140. Citeseer.

[Dekker, 2006] Dekker, S. (2006). The field guide to understanding human error.

[Demey et al., 2013] Demey, L., Kooi, B., and Sack, J. (2013). Logic and probability.

[Dimara et al., 2020] Dimara, E., Franconeri, S., Plaisant, C., Bezerianos, A., and Dragicevic, P. (2020). A task-based taxonomy of cognitive biases for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(2) :1413–1432.

[Dissing et al., 2020] Dissing, L. and Bolander, T. (2020). Implementing theory of mind on a robot using dynamic epistemic logic. In *IJCAI*, pages 1615–1621.

[Doherty et al., 1998] Doherty, P., Lukaszewicz, W., and Madalińska-Bugaj, E. (1998). The pma and relativizing minimal change for action update. volume 44, pages 258–269.

[Dupin De Saint-Cyr et al., 2014] Dupin De Saint-Cyr, F., Herzig, A., Lang, J., and Marquis, P. (2014). Raisonnement sur l'action et le changement. In *Panorama de l'intelligence artificielle, Volume 1 : représentation des connaissances et formalisation des raisonnements*, pages 363–392.

[Dupin de Saint-Cyr et al., 2011] Dupin de Saint-Cyr, F. and Lang, J. (2011). Belief extrapolation (or how to reason about observations and unpredicted change). *Artificial Intelligence*, 175(2) :760–790.

[Dutilh Novaes et al., 2016] Dutilh Novaes, C. and Veluwenkamp, H. (2016). Reasoning biases, non-monotonic logics and belief revision. *Theoria*, 83.

[Fermé et al., 2001] Fermé, E. L. and Hansson, S. O. (2001). *Shielded Contraction*, pages 85–107. Springer Netherlands, Dordrecht.

[Fermé et al., 2011] Fermé, E. L. and Hansson, S. O. (2011). Agm 25 years. *Journal of Philosophical Logic*, 40 :295–331.

[Fikes et al., 1971] Fikes, R. E. and Nilsson, N. J. (1971). Strips : A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4) :189–208.

[Finger, 1987] Finger, J. (1987). Exploiting constraints in design synthesis.

[Finucane et al., 2000] Finucane, M. L., Alhakami, A., Slovic, P., and Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of behavioral decision making*, 13(1) :1–17.

[Fouillard et al., 2021] Fouillard, V., Taha, S., Boulanger, F., and Sabouret, N. (2021). Belief revision theory. *Archive of Formal Proofs*. https://isa-afp.org/entries/Belief_Revision.html, Formal proof development.

[Frankish, 2010] Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10) :914–926.

[Frijda, 1988] Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, 43(5).

[Geffner, 1990] Geffner, H. (1990). Causal theories for nonmonotonic reasoning. In *AAAI*, pages 524–530.

[Gelfond et al., 1993] Gelfond, M. and Lifschitz, V. (1993). Representing action and change by logic programs. *The Journal of Logic Programming*, 17(2-4) :301–321.

[Gerbrandy et al., 1997] Gerbrandy, J. and Groeneveld, W. (1997). Reasoning about information change. *Journal of logic, language and information*, 6(2) :147–169.

[Gigerenzer et al., 2009] Gigerenzer, G. and Brighton, H. (2009). Homo heuristicus : Why biased minds make better inferences. *Topics in cognitive science*, 1(1) :107–143.

[Gärdenfors et al., 1995] Gärdenfors, P. and Rott, H. (1995). *Belief Revision*, volume 4, pages 35–132.

[Hanks et al., 1987] Hanks, S. and McDermott, D. (1987). Nonmonotonic logic and temporal projection. *Artificial intelligence*, 33(3) :379–412.

[Harel et al., 2001] Harel, D., Kozen, D., and Tiuryn, J. (2001). Dynamic logic. In *Handbook of philosophical logic*, pages 99–217. Springer.

[Harper, 1976] Harper, W. L. (1976). Rational conceptual change. In *PSA : Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1976, pages 462–494. Philosophy of Science Association.

[Herzig et al., 2010] Herzig, A., Lorini, E., Hübner, J. F., and Vercouter, L. (2010). A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1) :214–244.

[Herzig et al., 2017] Herzig, A., Lorini, E., Perrussel, L., and Xiao, Z. (2017). Bdi logics for bdi architectures : old problems, new perspectives. *KI-Künstliche Intelligenz*, 31(1) :73–83.

[Herzig et al., 1999] Herzig, A. and Rifi, O. (1999). Propositional belief base update and minimal change. *Artificial Intelligence*, 115(1) :107–138.

[Hinnosaar et al., 2012] Hinnosaar, M. and Hinnosaar, T. (2012). Authority bias.

[Hintikka, 1962] Hintikka, K. J. J. (1962). Knowledge and belief : An introduction to the logic of the two notions.

[Hójek, 2017] Hójek, A. (2017). Probability, logic, and probability logic. *The Blackwell guide to philosophical logic*, pages 362–384.

[Kahneman, 2003] Kahneman, D. (2003). Maps of bounded rationality : Psychology for behavioral economics. *American economic review*, 93(5) :1449–1475.

[Kahneman et al., 1982] Kahneman, D., Slovic, S. P., Slovic, P., and Tversky, A. (1982). *Judgment under uncertainty : Heuristics and biases*. Cambridge university press.

[Kahneman et al., 1979] Kahneman, D. and Tversky, A. (1979). Prospect theory : An analysis of decision under risk. *Econometrica*, 47(2) :263–292.

[Kaplan et al., 2016] Kaplan, R. L., Van Damme, I., Levine, L. J., and Loftus, E. F. (2016). Emotion and false memory. *Emotion Review*, 8(1) :8–13.

[Karau et al., 1993] Karau, S. J. and Williams, K. D. (1993). Social loafing : A meta-analytic review and theoretical integration. *Journal of personality and social psychology*, 65(4) :681.

[Katsuno et al., 1992] Katsuno, H. and Mendelzon, A. O. (1992). *On the difference between updating a knowledge base and revising it*, page 183–203. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

[Lazarus et al., 1984] Lazarus, R. S. and Folkman, S. (1984). *Stress, appraisal, and coping*. Springer publishing company.

[Lee et al., 2010] Lee, J. and Palla, R. (2010). Situation calculus as answer set programming. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[Lerner et al., 2015] Lerner, J. S., Li, Y., Valdesolo, P., and Kassam, K. S. (2015). Emotion and decision making. *Annual review of psychology*, 66(1).

[Leturc et al., 2018] Leturc, C. and Bonnet, G. (2018). A normal modal logic for trust in the sincerity. In *17th International Conference on Autonomous Agents and Multiagent Systems*.

[Levi, 1977] Levi, I. (1977). Subjunctives, dispositions and chances. *Synthese*, 34(4) :423–455.

[Liberatore et al., 1997] Liberatore, P. and Schaerf, M. (1997). Reducing belief revision to circumscription (and vice versa). *Artificial intelligence*, 93(1-2) :261–296.

[Liffiton et al., 2008] Liffiton, M. H. and Sakallah, K. A. (2008). Algorithms for computing minimal unsatisfiable subsets of constraints. *Journal of Automated Reasoning*, 40(1) :1–33.

[Lucas, 2001] Lucas, P. J. (2001). Bayesian model-based diagnosis. *International Journal of Approximate Reasoning*, 27(2) :99–119.

[Măirean et al., 2021] Măirean, C., Havârneanu, G. M., Barić, D., and Havârneanu, C. (2021). Cognitive biases, risk perception, and risky driving behaviour. *Sustainability*, 14(1) :77.

[Makinson, 1997] Makinson, D. (1997). Screened revision. *Theoria*, 63(1-2) :14–23.

[Marques-Silva et al., 2013] Marques-Silva, J., Heras, F., Janota, M., Previti, A., and Belov, A. (2013). On computing minimal correction subsets. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer.

[McCarthy, 1981] McCarthy, J. (1981). Epistemological problems of artificial intelligence. In *Readings in artificial intelligence*, pages 459–465. Elsevier.

[McCarthy, 1986] McCarthy, J. (1986). Applications of circumscription to formalizing common-sense knowledge. *Artificial intelligence*, 28(1) :89–116.

[McCarthy et al., 1969] McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press. reprinted in McC90.

[McDermott, 1982] McDermott, D. (1982). A temporal logic for reasoning about processes and plans. *Cognitive science*, 6(2) :101–155.

[Meyer et al., 2015] Meyer, J., Broersen, J., and Herzig, A. (2015). Bdi logics. *Handbook of Logics of Knowledge and Belief*, page 453.

[Milgram, 1963] Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4) :371.

[Moore, 1985] Moore, R. C. (1985). Semantical considerations on nonmonotonic logic. *Artificial intelligence*, 25(1) :75–94.

[Moura et al., 2008] Moura, L. d. and Bjørner, N. (2008). Z3 : An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer.

[Murata et al., 2015] Murata, A., Nakamura, T., and Karwowski, W. (2015). Influence of cognitive biases in distorting decision making and leading to critical unfavorable incidents. *Safety*, 1(1) :44–58.

[Narodytska et al., 2018] Narodytska, N., Bjørner, N., Marinescu, M. C., and Sagiv, M. (2018). Core-guided minimal correction set and core enumeration. In *IJCAI International Joint Conference on Artificial Intelligence : Stockholm, 13-19 July 2018*, pages 1353–1361. IJCAI.

[Nickerson, 1998] Nickerson, R. S. (1998). Confirmation bias : A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2) :175–220.

[Norling, 2004] Norling, E. (2004). Folk psychology for human modelling : Extending the bdi paradigm. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 202–209.

[Ortony et al., 1990] Ortony, A., Clore, G. L., and Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

[O'Sullivan et al., 2018] O'Sullivan, E. D. and Schofield, S. (2018). Cognitive bias in clinical medicine. *Journal of the Royal College of Physicians of Edinburgh*, 48(3) :225–232.

[Plaza, 1989] Plaza, J. (1989). Logics of public announcements. In *Proceedings 4th International Symposium on Methodologies for Intelligent Systems*.

[Plous, 1993] Plous, S. (1993). *The psychology of judgment and decision making.* Mcgraw-Hill Book Company.

[Poole, 1994] Poole, D. (1994). Representing diagnosis knowledge. *Annals of Mathematics and Artificial Intelligence*, 11(1) :33–50.

[Poole et al., 1987] Poole, D., Goebel, R., and Aleliunas, R. (1987). Theorist : A logical reasoning system for defaults and diagnosis. In *The Knowledge Frontier*, pages 331–352. Springer.

[Rao et al., 1991] Rao, A. S. and Georgeff, M. P. (1991). Modeling rational agents within a bdi-architecture. *KR*, 91 :473–484.

[Reason, 1990] Reason, J. (1990). The contribution of latent human failures to the breakdown of complex systems. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 327(1241) :475–484.

[Reiter, 1987] Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1) :57 – 95.

[Reiter, 1991] Reiter, R. (1991). *The Frame Problem in Situation the Calculus : A Simple Solution (Sometimes) and a Completeness Result for Goal Regression*, page 359–380. Academic Press Professional, Inc., USA.

[Salah et al., 2018] Salah, A. and Said Mehdi, D. (2018). L'entretien de recherche dit "semi-directif"dans les domaines des sciences humaines et sociales.

[Sandewall, 1995] Sandewall, E. (1995). *Features and fluents (vol. 1) the representation of knowledge about dynamical systems*. Oxford University Press, Inc.

[Sandewall, 1998] Sandewall, E. (1998). Cognitive robotics logic and its metatheory : Features and fluents revisited. *Electron. Trans. Artif. Intell.*, 2 :307–329.

[Schmitt-Beck, 2015] Schmitt-Beck, R. (2015). Bandwagon effect. *The international encyclopedia of political communication*, pages 1–5.

[Shanahan, 2016] Shanahan, M. (2016). The Frame Problem. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2016 edition.

[Sharot, 2011] Sharot, T. (2011). The optimism bias. *Current biology*, 21(23) :R941–R945.

[Shoham, 2009] Shoham, Y. (2009). Logical theories of intention and the database perspective. *Journal of Philosophical Logic*, 38(6) :633–647.

[Shoham, 2015] Shoham, Y. (2015). Why knowledge representation matters. *Communications of the ACM*, 59(1) :47–49.

[Simon, 1990] Simon, H. A. (1990). Bounded rationality. In *Utility and probability*, pages 15–18. Springer.

[Slater, 2018] Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British Journal of Psychology*, 109(3) :431–433.

[Slugoski et al., 1993] Slugoski, B. R., Shields, H. A., and Dawson, K. A. (1993). Relation of conditional reasoning to heuristic processing. *Personality and Social Psychology Bulletin*, 19(2) :158–166.

[Solaki et al., 2021] Solaki, A., Berto, F., and Smets, S. (2021). The logic of fast and slow thinking. *Erkenntnis*, 86(3) :733–762.

[Stanovich, 2009] Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds : Is it time for a tri-process theory ?

[Staw, 1997] Staw, B. M. (1997). The escalation of commitment : An update and appraisal.

[Takano et al., 1999] Takano, K. and Reason, J. (1999). Psychological biases affecting human cognitive performance in dynamic operational environments. *Journal of Nuclear Science and Technology*, 36(11) :1041–1051.

[Turner, 1999] Turner, H. (1999). A logic of universal causation. *Artificial Intelligence*, 113(1-2) :87–123.

[Tversky et al., 1971] Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, 76(2) :105.

[Tversky et al., 1974] Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty : Heuristics and biases. *Science*, 185(4157) :1124–1131.

[Tversky et al., 1983] Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning : The conjunction fallacy in probability judgment. *Psychological review*, 90(4) :293.

[Van Ditmarsch et al., 2011] Van Ditmarsch, H., Herzig, A., and De Lima, T. (2011). From situation calculus to dynamic epistemic logic. *Journal of Logic and Computation*, 21(2) :179–204.

[Van Ditmarsch et al., 2007] Van Ditmarsch, H. and Labuschagne, W. (2007). My beliefs about your beliefs : a case study in theory of mind and epistemic logic. *Synthese*, 155(2) :191–209.

[Van Linder et al., 1998] Van Linder, B., van der Hoek, W., and Meyer, J.-J. C. (1998). Formalising abilities and opportunities of agents. *Fundamenta Informaticae*, 34(1-2) :53–101.

[Voinson et al., 2015] Voinson, M., Billiard, S., and Alvergne, A. (2015). Beyond rational decision-making : modelling the influence of cognitive biases on the dynamics of vaccination coverage. *PloS one*, 10(11).

[Walmsley et al., 2016] Walmsley, S. and Gilbey, A. (2016). Cognitive biases in visual pilots' weather-related decision making. *Applied Cognitive Psychology*, 30(4) :532–543.

[Wassermann, 2000] Wassermann, R. (2000). An algorithm for belief revision. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning*, pages 345–352.

[Weaver et al., 2012] Weaver, E. A. and Stewart, T. R. (2012). Dimensions of judgment : Factor analysis of individual differences. *Journal of Behavioral Decision Making*, 25(4) :402–413.

[Wiegmann et al., 2017] Wiegmann, D. A. and Shappell, S. A. (2017). *A human error approach to aviation accident analysis : The human factors analysis and classification system*. Routledge.

[Woodcock et al., 2009] Woodcock, J., Larsen, P. G., Bicarregui, J., and Fitzgerald, J. (2009). Formal methods : Practice and experience. *ACM computing surveys (CSUR)*, 41(4) :1–36.

[Zhang et al., 2004] Zhang, J., Patel, V. L., Johnson, T. R., and Shortliffe, E. H. (2004). A cognitive taxonomy of medical errors. *Journal of biomedical informatics*, 37(3) :193–204.

Annexes

Annexe A : Équivalence avec un opérateur AGM

```
theory MCS

imports AGM_Revision

begin

context Supraclassical_logic

begin

text‹A remainder of @{term K} by @{term ϕ} is consistent›
theorem rem_consistent : ‹B ∈ K .⊥. ϕ ⟹ ¬ B ⊢ ⊥›
  by (metis (no_types, lifting) CollectD impI2 non_consistency notI_PL rem)

text‹If @{term B} is a remainder of @{term K} by false and @{term ϕ} is in it
 then @{term B} is also a remainder of @{term K} by ‹¬›@{term ϕ}›
theorem rem_bot:
  assumes as0: ‹ϕ ∈ B›
      and as1: ‹B ∈ K .⊥. ⊥›
    shows ‹B ∈ K .⊥. .¬ ϕ›
proof(unfold rem, intro  CollectI conjI allI impI)
  show ‹B ⊆ K› using as1 rem_inclusion by auto
next
  have ‹B ⊢ .¬ ϕ ⟹ B ⊢ ⊥›
    by (meson DiffI as0 as1 assumption_L non_consistency)
  then  show ‹¬ B ⊢ .¬ ϕ›
    by (meson as1 rem_consistent)
next
  fix B'
  assume as2:‹B' ⊆ K› and as3:‹B ⊂ B'›
  then have ‹B' ⊢ ⊥›
    using as1 rem by auto
  then show ‹B' ⊢ .¬ ϕ›
    by (meson impI2 notI_PL)
qed
```

```
text‹The set of Minimal Correction Set for @{term A} corresponds to all the subsets @{term B} of @{term A}
for which adding a subset of @{term B} in @{term A} makes @{term A} inconsistent›
definition MCSs::‹'a set ⇒ 'a set set›
  where ‹MCSs A ≡ {B. B ⊆ A ∧ ¬ (A - B ⊢ ⊥) ∧ (∀B' ⊂ B. A - B' ⊢ ⊥)}›

text‹@{term A} is consistent is equivalent that the set of MCS for @{term A} is a set of empty set›
corollary MCS_consistent_set_1: ‹¬ A ⊢ ⊥ ⟷ MCSs A = {{}}›
proof(rule iffI)
  show ‹¬ A ⊢ ⊥ ⟹ MCSs A = {{}}›
    unfolding MCSs_def psubsetI subset_antisym by auto
next
  show ‹MCSs A = {{}} ⟹ ¬ A ⊢ ⊥›
  unfolding MCSs_def
    by (metis (no_types, lifting) Diff_empty mem_Collect_eq singletonI)
qed

text‹@{term A} is consistent is equivalent that the set of MCS for @{term A} contains the empty set›
corollary MCS_consistent_set_2: ‹¬ A ⊢ ⊥ ⟷  {} ∈ MCSs A›
proof(rule iffI)
  show ‹¬ A ⊢ ⊥ ⟹ {} ∈ MCSs A›
    unfolding MCSs_def psubsetI subset_antisym by auto
next
  show ‹{} ∈ MCSs A ⟹ ¬ A ⊢ ⊥›
  unfolding MCSs_def by (metis (no_types, lifting) Diff_empty mem_Collect_eq)
qed

text‹If @{term M} is a MCS of @{term K} then ‹K - M› is a remainder of @{term K} by false›
lemma MCS_remBot: ‹M ∈ MCSs K ⟹ K - M ∈ K .⊥. ⊥›
proof(unfold rem MCSs_def, elim CollectE conjE, intro impI conjI CollectI allI, simp_all, goal_cases)
  case (1 B)
  then show ?case
    using 1(3)[rule_format, of ‹K - B›]
    by (metis Diff_Diff_Int Diff_mono equalityD1 inf.absorb_iff2 leD psubsetI psubset_imp_subset)
qed
```

```
text‹If @{term B} is a remainder of @{term K} by false then ‹K - B› is a MCS of @{term K}›
lemma remBot_MCS: ‹B ∈ K .⊥. ⊥ ⟹ K - B ∈ MCSs K›
proof(unfold rem MCSs_def, elim CollectE conjE, intro impI conjI CollectI allI, simp_all, goal_cases)
  case 1
  then show ?case
    by (metis Diff_Diff_Int inf.absorb_iff2)
next
  case (2 B')
  then show ?case
    using 2(3)[rule_format, of ‹K - B'›] by blast
qed

text‹▪ To say @{term M} is a MCS of @{term K} is equivalent to say that @{term M} is a subset of @{term K}
and ‹K - M› is a remainder of @{term K} by false
▪ To say  @{term B} is a remainder of @{term K} by false is equivalent to say that @{term B} is a subset of @{term K}
and ‹K - B› is a MCS of @{term K}›
theorem MCS_remBot_unfold: ‹M ∈ MCSs K ⟷ (M ⊆ K ∧ K - M ∈ K .⊥. ⊥)›
    and remBot_MCS_unfold: ‹B ∈ K .⊥. ⊥ ⟷ (B ⊆ K ∧ K - B ∈ MCSs K)›
  by (metis (no_types, lifting) Diff_Diff_Int MCS_remBot MCSs_def inf.absorb_iff2 mem_Collect_eq remBot_MCS)
     (metis Diff_Diff_Int MCS_remBot inf.absorb_iff2 remBot_MCS rem_inclusion)

text‹If @{term M} is a MCS of @{term K} and @{term ϕ} is in ‹K - M› then ‹K - M› corresponds to
a remainder of @{term K} by ‹¬›@{term ϕ}›
theorem MCS_remain: ‹M ∈ MCSs K ⟹ ϕ ∈ K - M ⟹ K - M ∈ K .⊥. (.¬ ϕ)›
  using MCS_remBot_unfold rem_bot by meson
```

```
text‹If @{term M} is a MCS of ‹K ∪ ¬ϕ› and @{term M} doesn't contain ‹¬›@{term ϕ} then
‹K - M› is a remainder of @{term K} by @{term ϕ}›
lemma MCS_rem:
  assumes a1:‹M ∈ MCSs (K ∪ {.¬ ϕ})› and a2:‹.¬ ϕ ∉ M›
  shows ‹K - M ∈ K .⊥. ϕ›
proof(cases ‹.¬ ϕ ∈ K›)
  case True
  with a1 a2 show ?thesis
    using MCS_remain by (metis Cn_notnot Diff_iff Un_insert_right insert_absorb remainder_extensionality sup_bot.right_neutral)
next
  case False
  with a1 a2 show ?thesis
  proof(unfold rem MCSs_def, safe, goal_cases)
    case 1
    then show ?case
      by (metis Un_insert_right insert_Diff_if not_PL notnot_PL sup_bot.right_neutral)
  next
    case (2 B' ψ)
    then show ?case by (meson subsetD)
  next
    case (3 B' ψ)
    have a3:‹K ∪ {.¬ ϕ} - (M - {ψ}) = (K - (M - {ψ})) ∪ {.¬ ϕ}›
      using a2 by fastforce
    have a4:‹K - (M - {ψ}) ⊢ ϕ›
      using 3(8)[rule_format, of ‹M - {ψ}›, simplified a3] 3
      by (metis Diff_iff Diff_subset absurd_PL insertCI not_PL psubsetI)
    have a5:‹K - (M - {ψ}) ⊆ B'›
       using 3 by blast
    have ‹B' ⊢ ϕ›
      by (meson a5 a4 assumption_L subset_eq transitivity2_L)
    with 3 show ?case
      by blast
  qed
qed
```

```
text‹If @{term B} is a remainder of @{term K} by @{term ϕ} then ‹K - B› is a MCS of ‹K ∪ ¬ϕ››
lemma rem_MCS:
  assumes a1: ‹B ∈ K .⊥. ϕ›
  shows ‹K - B ∈ MCSs (K ∪ {.¬ ϕ})›
using a1 unfolding rem MCSs_def
proof(elim CollectE conjE, intro impI conjI CollectI allI, goal_cases)
  case 1
  then show ?case by blast
next
  case 2
  then show ?case
    by (metis Un_insert_right absurd_PL assumption_L double_diff insertCI insert_Diff_if non_consistency
          not_PL subset_refl sup_bot.right_neutral)
next
  case (3 B')
  have a3:‹K - B' ⊢ ϕ›
    using 3(3)[rule_format, of ‹K - B'›] 3 by blast
  with 3 show ?case
  proof(cases ‹.¬ ϕ ∈ B'›)
    case True
    have ‹B ∪ {.¬ ϕ} ⊢ ϕ›
      using 3(3)[rule_format, of ‹B ∪ {.¬ ϕ}›] 3 True by blast
      with True 3 show ?thesis
        by (metis (no_types, lifting) Cn_not notnot_PL sup.right_idem)
  next
    case False
    with a3 3 show ?thesis by (metis Un_empty_right Un_insert_right insert_Diff_if not_PL notnot_PL)
  qed
qed

text‹If @{term B} is a remainder of @{term K} by @{term ϕ} then ‹¬›@{term ϕ} is not in the MCS of @{term K}›
lemma rem_MCS_sound:
 assumes a1: ‹B ∈ K .⊥. ϕ›
  shows ‹.¬ ϕ ∉ K - B›
  using a1 unfolding rem
proof(safe, goal_cases)
  case 1
  have ‹B ∪ {.¬ ϕ} ⊢ ϕ›
    using 1(3)[rule_format, of ‹B ∪ {.¬ ϕ}›] 1 by blast
  hence ‹B ⊢ ϕ›
    by (metis not_PL notnot_PL sup.right_idem)
  with 1(2) show ?case by rule
qed
```

```
text‹▪ To say that ‹¬›@{term ϕ} is not in the MCS @{term M} of ‹K ∪ ¬ϕ› is equivalent to say that @{term M} is a subset of @{term K}
and ‹K - M› is a remainder of @{term K} by @{term ϕ}
▪ To say that @{term B} is a remainder of @{term K} by @{term ϕ} is equivalent to say that @{term B} is a subset of @{term K} and ‹K - B› is a MCS of ‹K ∪ ¬ϕ›
where ‹¬›@{term ϕ} is not in ‹K - B››
theorem MCS_rem_unfold: ‹.¬ ϕ ∉ M ⟹ M ∈ MCSs (K ∪ {.¬ ϕ}) ⟷ (M ⊆ K ∧ K - M ∈ K .⊥. ϕ)›
    and rem_MCS_unfold: ‹B ∈ K .⊥. ϕ ⟷ (B ⊆ K ∧ K - B ∈ MCSs (K ∪ {.¬ ϕ}) ∧ .¬ ϕ ∉ K - B)›
  by (metis Diff_Diff_Int MCS_rem MCS_remBot_unfold Un_insert_right inf.absorb_iff2 rem_MCS subset_insert sup_bot_right)
     (metis MCS_rem double_diff rem_MCS rem_MCS_sound rem_inclusion subset_refl)

text‹A remainder of @{term K} by @{term ϕ} is a MCS @{term M} of @{term K} that doesn't contain ‹¬›@{term ϕ}›
theorem MCS_rem_graal: ‹K .⊥. ϕ = {K - M|M. M ∈ MCSs (K ∪ {.¬ ϕ}) ∧ .¬ ϕ ∉ M}›
proof(safe)
  fix B
  assume ‹B ∈ K .⊥. ϕ›
  then show ‹∃M. B = K - M ∧ M ∈ MCSs (K ∪ {.¬ ϕ}) ∧ .¬ ϕ ∉ M›
    using rem_MCS_unfold by auto
next
  fix M
  assume ‹M ∈ MCSs (K ∪ {.¬ ϕ})› and ‹.¬ ϕ ∉ M›
  then show ‹K - M ∈ K .⊥. ϕ›
    using MCS_rem_unfold by blast
qed
```

```
text‹In a finite set @{term K} and if @{term ϕ} is not a tautologie there exists a MCS of @{term K} that doesn't contain ‹¬›@{term ϕ}›
lemma MCS_exists_selection_finite: ‹finite K ⟹ ¬ ⊩ ϕ ⟹ ∃M. M ∈ MCSs K ∧ .¬ ϕ ∉ M›
  unfolding MCS_remBot_unfold rem
proof (simp del:infer_def, case_tac ‹.¬ ϕ ∈ K›, goal_cases)
  case 1
  define A where a0:‹A = {B. B ⊆ K ∧ ¬ B ⊢ ⊥ ∧ .¬ ϕ ∈ B}›
  hence a1:‹finite A› by (simp add: 1)
  have a2:‹{.¬ ϕ} ∈ A›
    using 1 a0 notnot_PL valid_def valid_not_PL by fastforce
  obtain B where a3:‹B ∈ A ∧ (∀B'∈A. B ⊆ B' ⟶ B = B')›
    by (metis a1 a2 empty_iff finite_has_maximal)
  hence a4:‹K - (K - B) = B›
    using a0 by blast
  show ?case
    apply(rule_tac x=‹K - B› in exI, subst (1 2) a4)
    using a0 a3 by blast
next
  case 2
  define A where a0:‹A = {B. B ⊆ K ∧ ¬ B ⊢ ⊥}›
  hence a1:‹finite A› by (simp add: 2)
  have ‹¬ ⊩ ⊥›
    by (metis 2(2) Cn_true UNIV_I infer_def valid_Cn_not valid_def valid_not_PL)
  hence a2:‹{} ∈ A›
    by (simp add: a0 valid_def)
  obtain B where a3:‹B ∈ A ∧ (∀B'∈A. B ⊆ B' ⟶ B = B')›
    by (metis a1 a2 empty_iff finite_has_maximal)
  hence a4:‹K - (K - B) = B›
    using a0 by blast
  show ?case
    apply(rule_tac x=‹K - B› in exI, subst (1 2) a4)
    using 2 a0 a3 by blast
qed
```

```
locale MCS_Correction = Supraclassical_logic + Compact_logic

begin

text‹A selection function ‹γMCS› for @{term K} and @{term ϕ} return @{term K} if @{term ϕ} is valid
else return ‹K - M› where @{term M} is a MCS of ‹K ∪ ¬ϕ› that doesn't contain ‹¬›@{term ϕ}›
definition MCS_selection (‹γMCS›)
  where ‹γMCS K ϕ = if ⊩ ϕ then {K}
                          else {K - M|M. M ∈ MCSs (K ∪ {.¬ ϕ}) ∧ .¬ ϕ ∉ M}›

text‹‹γMCS› corresponds to a full meet contraction›
sublocale FullMeetContraction where full_sel = ‹γMCS›
  apply (unfold_locales)
  unfolding MCS_selection_def emptyremtaut MCS_rem_graal[symmetric] by assumption

text‹‹γMCS› corresponds to a transitively relational meet contraction›
sublocale TRMC_SC where relation = ‹λ K A B. True› and rel_sel =  ‹γMCS›
  apply (unfold_locales, simp_all)
  using remainder_extensionality by blast

text‹Using ‹γMCS› to contract corresponds to an AGM full contraction›
sublocale AGMFC_S where contraction = ‹λA ϕ. A  ÷⇘γMCS⇙ ϕ›
  by (unfold_locales)

text‹Using ‹γMCS› to contract corresponds to an AGM full revision›
sublocale AGM_FullRevision where revision = ‹λK ϕ. (K ÷⇘γMCS⇙ .¬ ϕ) ⊕ ϕ›
  by  (unfold_locales)

text‹All AGM axioms are satisfied for the contraction and revision throught the MCS›
lemmas MCS_lemmas = contract_closure contract_inclusion contract_vacuity contract_success contract_recovery contract_extensionality
                    contract_conj_overlap contract_conj_inclusion
                    revis_closure revis_inclusion revis_vacuity revis_success revis_extensionality revis_consistency
                    revis_superexpansion revis_subexpansion
                    is_selection tautology_selection nonempty_selection extensional_selection
```

```
text‹If @{term ϕ} is not a tautologie then there exists a MCS of @{term K} that doesn't contain ‹¬›@{term ϕ}
(required Compact Logic in the infinite case)›
lemma MCS_exists_selection: ‹¬ ⊩ ϕ ⟹ ∃M. M ∈ MCSs K ∧ .¬ ϕ ∉ M›
  unfolding MCS_remBot_unfold
proof (case_tac ‹.¬ ϕ ∈ K›, goal_cases)
  case 1
  then show ?case
    using upper_remainder[of ‹{.¬ ϕ}› K ⊥]
    by (metis Diff_iff Diff_subset double_diff empty_subsetI insert_subset notnot_PL rem_inclusion subsetI valid_def valid_not_PL)
next
  case 2
  then show ?case
    using upper_remainder[of ‹{}› K ⊥]
    by (metis Diff_Diff_Int Diff_iff empty_subsetI inf.absorb_iff2 non_consistency not_PL rem_inclusion transitivity2_L valid_def)
qed

lemma MCS_exists_selection_v2: ‹¬ ⊩ (.¬ ϕ) ⟹ ∃M. M ∈ MCSs K ∧ ϕ ∉ M›
  unfolding MCS_remBot_unfold
proof (case_tac ‹ϕ ∈ K›, goal_cases)
  case 1
  then show ?case
    using upper_remainder[of ‹{ϕ}› K ⊥]
    by (metis Diff_iff MCS_remBot_unfold empty_subsetI insert_subset remBot_MCS valid_not_PL)
next
  case 2
  then show ?case
    using upper_remainder[of ‹{}› K ⊥]
    by (meson MCS_exists_selection MCS_remBot_unfold in_mono)
qed

lemma MCS_exists: ‹¬ ⊩ ⊥ ⟹ ∃M. M ∈ MCSs K›
  unfolding MCS_remBot_unfold
  using upper_remainder[of ‹{}› K ⊥]
  by (meson Diff_subset MCS_remBot empty_subsetI remBot_MCS valid_def)
```

Annexe B : Complétude et correction de l'algorithme

```
— ‹MCS incremental›

text‹If @{term A} is a subset of @{term B} and @{term M₁} is a MCS of @{term A}
and @{term M₂} a MCS of ‹B - M₁› then there exists a MCS of @{term B} that correspond to ‹ M₁ ∪ M₂››
theorem MCS_steps_soundness_weak:
  assumes inc:‹A ⊆ B› and step1: ‹M₁ ∈ MCSs A› and step2:‹M₂ ∈ MCSs (B - M₁)›
    shows ‹∃M∈MCSs B. M ⊆ M₁ ∪ M₂›
proof -
  have a0:‹B - M₁ - M₂ ⊆ B› by blast
  have a1:‹¬ B - M₁ - M₂ ⊢ ⊥› using MCSs_def step2 by blast
  show ?thesis
    using upper_remainder[OF a0 a1]
    by (metis Diff_subset_conv remBot_MCS sup_commute sup_left_commute)
qed

text‹If @{term A} is a subset of @{term B} and @{term M₁} is a MCS of @{term A}
and @{term M₂} a MCS of ‹B - M₁› then there exists a MCS @{term M} of @{term B} that correspond to ‹ M₁ ∪ M₂›
and @{term M₂} is a subset of @{term M}›
theorem MCS_steps_soundness_strong: — ‹a way to improve our algorithm›
  assumes inc:‹A ⊆ B› and step1: ‹M₁ ∈ MCSs A› and step2:‹M₂ ∈ MCSs (B - M₁)›
    shows ‹∃M∈MCSs B. M₂ ⊆ M ∧ M ⊆ M₁ ∪ M₂›
proof -
  obtain M where a0:‹M ∈ MCSs B ∧  M ⊆ M₁ ∪ M₂›
    using MCS_steps_soundness_weak inc step1 step2 by blast
  { fix ϕ
    assume b0:‹ϕ ∈ M₂› and b1:‹ϕ ∉ M›
    hence b1:‹(B - M₁) - (M₂ - {ϕ}) ⊆ B - M›
      using inc a0 by auto
    have ‹(B - M₁) - (M₂ - {ϕ}) ⊢ ⊥›
      using step2[simplified MCSs_def]
      using b0 double_diff by auto
    with b1 have ‹B - M ⊢ ⊥›
      by (meson assumption_L subsetD transitivity2_L)
  } note a1=this
  show ?thesis
    by (meson MCS_remBot a0 a1 rem_consistent subsetI)
qed
```

```
text‹If @{term A} is a subset of @{term B} and @{term M₁} is a MCS of @{term A}
and @{term M₂} a MCS of ‹B - M₁› and if @{term M₂} doesn't contain proposition from @{term A} then ‹M₁ ∪ M₂› is minimal for @{term B}›
theorem MCS_steps_match1:
  assumes inc:‹A ⊆ B› and step1: ‹M₁ ∈ MCSs A› and step2:‹M₂ ∈ MCSs (B - M₁)›
      and MCS_inc:‹A ∩ M₂ = {}›
    shows ‹M₁ ∪ M₂ ∈ MCSs B›
  unfolding MCSs_def
proof(simp del:infer_def, intro conjI allI impI, goal_cases)
  case 1
  then show ?case
    using inc step1[simplified MCSs_def] by fastforce
next
  case 2
  then show ?case
    using inc step2[simplified MCSs_def] by blast
next
  case 3
  then show ?case
    using step2[simplified MCSs_def] by simp (metis Diff_Diff_Int Diff_Un Diff_subset double_diff)
next
  case (4 B')
  have a0:‹M₁ ∩ M₂ = {}› and a1:‹M₂ ⊆ B - A› and a2:‹M₁ ⊆ A›
    using step1 step2 MCSs_def MCS_inc by auto
  { fix ϕ
    assume b0:‹ϕ ∈ M₂› and b1:‹ϕ ∉ B'›
    hence b1:‹(B - M₁) - (M₂ - {ϕ}) ⊆ B - B'›
      using 4 inc by auto
    have ‹(B - M₁) - (M₂ - {ϕ}) ⊢ ⊥›
      using step2[simplified MCSs_def]
      using b0 double_diff by auto
    with b1 have ‹B - B' ⊢ ⊥›
      by (meson assumption_L subsetD transitivity2_L)
  } note a3=this
  { fix ϕ
    assume b0:‹ϕ ∈ M₁› and b1:‹ϕ ∉ B'›
    hence b1:‹A - (M₁ - {ϕ}) ⊆ B - B'›
      using 4 inc MCS_inc by auto
    have ‹A - (M₁ - {ϕ}) ⊢ ⊥›
      using step1[simplified MCSs_def]
      using b0 double_diff by auto
    with b1 have ‹B - B' ⊢ ⊥›
      by (meson assumption_L subsetD transitivity2_L)
  } note a4=this
  show ?case
    using 4 a3 a4 by auto
qed
```

```
text‹If @{term A} is a subset of @{term B} and @{term M₁} is a MCS of @{term A} and ‹B -  M₁› is consistent then
@{term M₁} is a MCS of @{term B}›
corollary MCS_steps_match2:
  assumes inc:‹A ⊆ B› and step1: ‹M₁ ∈ MCSs A› and step2:‹¬ B - M₁ ⊢ ⊥› — ‹no need for a compact logic›
    shows ‹M₁ ∈ MCSs B›
  using inc step1 apply(rule MCS_steps_match1[where M₂=‹{}›, simplified])
  using MCS_consistent_set_1 step2 by blast

text‹If @{term A} is a subset of @{term B} then there exists a MCS @{term M₁} of @{term A}
and a MCS @{term M₂} of @{term B} where ‹M₁ ∪ M₂› is a MCS of only @{term B} ›
theorem MCS_steps_completeness:
  assumes inc:‹A ⊆ B› and mcs: ‹M ∈ MCSs B›
    shows ‹∃M₁ M₂. M₁ ∈ MCSs A ∧ M₂ ∈ MCSs (B - M₁) ∧ M = M₁ ∪ M₂›
proof -
  have a0:‹¬ A - M ⊢ ⊥›
    using monotonicity_L[of ‹A - M› ‹B - M›]
    by (metis MCS_remBot_unfold Un_Diff inc infer_def mcs rem_consistent subset_iff sup.absorb_iff1)
  obtain M₁ where a1:‹M₁ ∈ MCSs A ∧ M₁ ⊆ M›
    unfolding MCS_remBot_unfold
    using upper_remainder[of ‹A - M› A ⊥, OF _ a0]
    by (metis Diff_subset Diff_subset_conv MCS_remBot Un_commute remBot_MCS)
  have a2:‹¬ B - M₁ - M ⊢ ⊥›
    using monotonicity_L[of ‹B - M₁ - M› ‹B - M›]
    using MCS_remBot mcs rem_consistent by fastforce
  obtain M₂ where a3:‹M₂ ∈ MCSs (B - M₁) ∧ M₂ ⊆ M›
    unfolding MCS_remBot_unfold
    using upper_remainder[of ‹B - M₁ - M› ‹B - M₁› ⊥, OF _ a2, simplified]
    by (metis Diff_subset_conv MCS_remBot_unfold Un_commute remBot_MCS_unfold)
  have a4:‹M₁ ∪ M₂ ⊆ M›
    by (simp add: a1 a3)
  obtain M' where a5:‹M' ∈ MCSs B ∧ M' ⊆ M₁ ∪ M₂›
    by (meson MCS_steps_soundness_weak a1 a3 inc)
  hence a6:‹M' = M›
    by (metis (no_types, lifting) CollectD MCSs_def a4 dual_order.strict_trans1 mcs psubsetI)
  show ?thesis
    using a1 a3 a4 a5 a6 by blast
qed
```