Capturer les biais cognitifs dans un processus de prise de décision erroné

Valentin Fouillard, Nicolas Sabouret, Safouan Taha, Frédéric Boulanger

▶ **To cite this version:**

Valentin Fouillard, Nicolas Sabouret, Safouan Taha, Frédéric Boulanger. Capturer les biais cognitifs dans un processus de prise de décision erroné. WACAI 2021, Centre National de la Recherche Scientifique [CNRS], Oct 2021, Saint Pierre d'Oléron, France. hal-03377541

HAL Id: hal-03377541
https://hal.science/hal-03377541

Submitted on 14 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Capturer les biais cognitifs dans un processus de prise de décision erroné

Valentin Fouillard
Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique
91405, Orsay, France

Nicolas Sabouret
Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique
91405, Orsay, France

Safouan Taha
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire méthodes formelles
91190, Gif-sur-Yvette, France

Frédéric Boulanger
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire méthodes formelles
91190, Gif-sur-Yvette, France

RÉSUMÉ

Cet article présente un modèle logique pour étudier le raisonnement d'un opérateur humain et son altération par des biais cognitifs dans des situations d'accident. Ce modèle formel de diagnostic s'appuie sur des révisions de croyance minimales pour reconstituer le comportement apparemment irrationnel de l'opérateur. Nous appliquons ce modèle à l'accident du vol 447 d'Air France, et nous montrons qu'un tel modèle permet de retrouver des motifs de biais cognitifs soupçonnés d'avoir joué un rôle dans le comportement des pilotes ainsi que d'autres motifs de biais plausibles.

KEYWORDS

Biais cognitifs, révision de croyance, diagnostic, prise de décision

1 INTRODUCTION

Trois ans après le crash du vol 447 Rio-Paris d'Air France en juillet 2012, le rapport du Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile (BEA) est publié [1]. Cette enquête montre que les pilotes n'ont pas identifié le décrochage de l'avion alors que l'alarme de décrochage a retenti plus de 75 fois. D'un point de vue extérieur, le comportement des pilotes peut sembler totalement irrationnel. Néanmoins, le BEA souligne une possible confusion entre une situation de décrochage et de survitesse. Plusieurs éléments soutiennent cette hypothèse : le manque d'information visuelle, de mauvaises indications du directeur de vol, une alarme de décrochage irrégulière, etc. Si nous faisons l'hypothèse que les pilotes pensaient être en survitesse alors toutes leurs actions ont été rationnelles.

Dans cet article, nous nous intéressons à l'étude, à l'aide de la simulation informatique, de telles situations où un opérateur humain peut adopter un comportement erroné, irrationnel au regard de l'ensemble des informations qui lui ont été transmises, mais pourtant cohérent avec une forme de raisonnement. Notre objectif est de déterminer les causes possible des erreurs en nous appuyant sur les états mentaux possibles de l'opérateur, reconstruits à partir de ses actions et observations (section 2). Pour cela, nous définissons un modèle basé sur la logique formelle (section 3). En nous appuyant sur les mécanismes de révision de croyance en logique, nous générons l'ensemble des états possibles. Nous proposons ensuite un mécanisme de détection de biais cognitifs (section 4) pour identifier les erreurs plausibles sur le plan psychologique. Nous discutons des travaux connexes dans la section 5 et nous concluons par discuter des possibilités offertes par ce modèle (section 6).

2 DESCRIPTION DU PROBLÈME

Nous souhaitons modéliser une situation dans laquelle un opérateur adopte un comportement irrationnel. Par comportement irrationnel, nous entendons une action (ou un ensemble d'actions) en contradiction avec l'état du monde. Par exemple dans le cas du crash du vol 447, nous pouvons observer que le pilote a effectué une série d'actions maintenant l'appareil en situation de décrochage au lieu de sortir de cette situation. Pour cela, nous disposons (voir Figure 1) :

- d'une « trace », c'est-à-dire une suite d'actions effectuées par l'opérateur au cours du temps ;
- de toutes les observations que l'opérateur peut effectuer à chaque pas de temps ;
- des croyances initiales B_0 de l'opérateur ;
- les règles du raisonnement de l'opérateur.

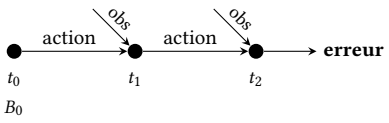


Figure 1: Croyances initiales, actions et observations

Nous définissons un comportement rationnel comme une suite d'états de croyance, cohérente avec les règles de raisonnements, les actions et les observations effectuées à chaque instant. Deux problèmes se posent :

- Certaines nouvelles observations peuvent être incompatibles avec les croyances ou les déductions de l'opérateur (*e.g.* j'observe des nuages alors que je croyais qu'il n'y a pas de nuages) ;
- Les croyances ou les déductions de l'opérateur, tenant compte de l'ensemble des observations et croyances passées, sont incompatibles avec l'action effectuée (c'est la situation de comportement irrationnel que nous voulons modéliser).

Notre problème est alors de rétablir la cohérence des croyances de l'opérateur à chaque instant. Nous devons pour cela ignorer certaines croyances et observations de l'opérateur, de manière à ce que les croyances conservées, les observations retenues et les actions

effectuées soient cohérentes, donc conformes au comportement rationnel de l'opérateur que nous voulons modéliser. Ce problème est bien connu en informatique et en logique formelle : il s'agit du problème de la *révision de croyances* [11]. Nous prétendons que chaque révision correspond à un état mental possible de l'opérateur qui pourrait expliquer sa décision erronée. Dans l'exemple du vol AF-447, ignorer l'alarme peut conduire les pilotes à croire qu'ils sont en situation de survitesse, ce qui explique leurs actions.

Notre objectif est alors de décider quelles révisions de croyances sont acceptables d'un point de vue psychologique.

3 MODÈLE

3.1 Nomenclature

Nous basons notre modèle sur une logique de prédicat du premier ordre où chaque prédicat est indicé temporellement afin d'exprimer l'instant auquel est considéré le prédicat. Par exemple :

$\text{nuages}(x,y)_t \quad \rightarrow \quad$ il y a des nuages en (x,y) au temps t
$\text{alarme}_t \quad \rightarrow \quad$ l'alarme sonne au temps t

Pour simplifier l'écriture des états, toute variable ou indice temporel libre est considéré implicitement comme quantifié universellement :

$$P(x,y)_t \quad \Leftrightarrow \quad \forall x, \forall y, \forall t \quad P(x,y)_t$$
$$P(x,2)_1 \quad \Leftrightarrow \quad \forall x \quad P(x,2)_1$$

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

$$\alpha ::= p \mid \neg\alpha \mid \alpha_1 \wedge \alpha_2 \mid \alpha_1 \vee \alpha_2 \mid \alpha_1 \Rightarrow \alpha_2$$

où α est une formule valide de $\mathcal{L}_0$ et $p \in \mathcal{P}red$, l'ensemble des prédicats indicés temporellement. La description des croyances de l'opérateur et des observations, à chaque instant, ainsi que des règles de raisonnement, se fait à l'aide de formules de $\mathcal{L}_0$. Les règles de raisonnement sont numérotées. Ainsi la règle suivante indique qu'à tout instant, en toute position, s'il n'y a pas de nuages et qu'il y a du soleil, le ciel est bleu :

$$R_1(x)_t \equiv ((\neg nuages(x)_t \wedge soleil(x)_t) \Rightarrow cielBleu(x)_t)$$

On peut ainsi « ignorer » une croyance, une observation ou une règle lors de la révision de croyance. Sur cet exemple, ignorer R_1 signifie qu'on ne peut plus déduire que le ciel est bleu lorsqu'il y a du soleil et pas de nuages.

Nous définissons aussi un ensemble de symboles indicés par le temps appelé $\mathcal{A}$ qui représente les 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 $\mathcal{A}$ à l'ensemble des prédicats, ainsi que les deux opérateurs suivants :

$$\phi ::= \alpha \mid [\alpha]act \mid act :: \alpha$$

avec $\alpha \in \mathcal{L}_0$ et $act \in \mathcal{A}$. $[\alpha]act$ exprime que α est la précondition de l'action act et $act :: \alpha$ exprime que α est l'effet de l'action act. Par exemple :

$[\neg \text{locked}_t]\,\text{doOpen}_t$ — l'action doOpen nécessite que la porte ne soit pas verrouillée.

$\text{doOpen}_t :: \text{open}_{t+1}$ — l'action open a pour effet que la porte est ouverte au pas de temps suivant.

Le langage $\mathcal{L}$ nous permet d'écrire un ensemble de propositions caractérisant un comportement apparemment irrationnel.

3.2 Description d'un comportement

La description d'un comportement irrationnel passe par la définition de plusieurs éléments :

Les croyances initiales $\mathcal{B}elief_{init}$: ensemble de prédicats $p \in \mathcal{P}red$ représentant les croyances initiales de l'opérateur sur le monde.

Les règles de raisonnement $\mathcal{R}$: ensemble de formules du langage $\mathcal{L}$ avec un indice temporel t libre, donc applicables à tout instant t. Certaines de ces règles permettent de déduire de nouvelles propositions à partir de croyances et d'observations (*e.g* $(\neg nuages(x)_t \wedge soleil(x)_t) \Rightarrow cielBleu(x)_t$). D'autres définissent les préconditions et les effets des actions (*e.g* $[\neg \text{locked}_t]\,\text{doOpen}_t$).

Les désirs $\mathcal{D}$: ensemble de littéraux positifs ou négatifs de $\mathcal{P}red$. Ils représentent les buts de l'opérateur, c'est à dire les croyances qui doivent être satisfaites au *prochain* instant.

Les observations indicées dans le temps $Obs = \{Obs_1, ..., Obs_t\}$: chaque ensemble d'observations Obs_i est un ensemble de littéraux positifs $o \in \mathcal{P}red$ ou négatifs $\neg o \in \mathcal{P}red$. Les littéraux qui ne sont pas renseignés dans une observation sont inconnus pour cet instant (ce qui est différent d'observer la négation du prédicat).

La trace $\mathcal{T} = \{a_1, ..., a_t\}$: ensemble d'actions de $\mathcal{A}$ représentant les actions que l'opérateur a effectuées à chaque pas de temps.

Les règles de cohérence C : ensemble de formules propositionnelles vraies en tout temps et qui ne peuvent pas être « ignorées » dans le processus de révision de croyance. Ces règles permettent de décrire des propriétés physiques (par exemple pour dire que deux actions sont incompatibles : $(avancer_t \wedge attendre_t) \Rightarrow \bot$).

À partir de ces ensembles, notre objectif est de calculer à chaque instant les états mentaux possibles de l'opérateur, c'est-à-dire un ensemble de propositions représentant les croyances de l'opérateur à un instant t. La prochaine section définit formellement comment l'état mental à l'instant t est calculé à partir des états mentaux précédents, des observations et de la trace d'actions.

3.3 Définition d'un état mental

Un état mental est un ensemble de prédicats et de règles dans le langage $\mathcal{L}$ décrivant des croyances, des désirs, des observations et des règles de raisonnement qui ont été *retenues* par l'agent (certaines décrivent des préconditions et des effets d'actions), ainsi qu'une trace d'actions effectuées par le passé. Nous notons B_t l'état mental à l'instant t. Nous définissons l'état mental initial de l'agent comme la conjonction des croyances initiales, des règles de raisonnement, des désirs (appliqués à l'instant suivant) et des règles de cohérence :

$$B_0 = \mathcal{B}elief_{init} \wedge \mathcal{R} \wedge \mathcal{D} \wedge C$$

À chaque instant, B_t ne contient qu'un sous-ensemble de $\mathcal{L}$: tout ce qui n'est pas dans B_t est inconnu de l'agent. Ainsi :

$\text{p}_{t'} \in B_t \quad \rightarrow$ Je crois à l'instant t que p est **vrai** à l'instant t'

$\neg\,\text{p}_{t'} \in B_t \quad \rightarrow$ Je crois à l'instant t que p est **faux** à l'instant t'

$\text{p}_{t'} \notin B_t \wedge \neg\,\text{p}_{t'} \notin B_t \quad \rightarrow$ Je **ne sais pas** à l'instant t si p est vrai ou faux à l'instant t'

Afin que l'opérateur se comporte de manière rationnelle, notre modèle impose que chaque B_t soit consistant : l'ensemble des croyances de l'agent est cohérent avec ses actions, ses observations, ses désirs et ses règles de raisonnement. Nous devons alors déterminer comment le modèle transite d'un état mental B_{t-1} (consistant) vers un nouvel état mental B_t (consistant) en tenant compte des observations Obs_t et de l'action a_t, qui peuvent justement créer des inconsistances. C'est l'objet de la *révision de croyances*.

Notre première hypothèse est que *l'opérateur continue de croire ce qu'il croyait au temps précédent* : chaque B_t contient a priori l'ensemble des éléments de l'état mental précédent B_{t-1}.

Dans un premier temps, nous ajoutons les observations à la base de croyances de l'opérateur. Si ces observations ne remettent pas en cause ses croyances, c'est-à-dire si l'ensemble obtenu $B_{t-1} \cup Obs_t$ est consistant, alors il n'y a pas de problème. Sinon, il faut faire une révision de croyances pour rester cohérent.

Dans un deuxième temps, puisque nous savons que l'opérateur a effectué l'action a_t au pas de temps t, nous devons nous assurer que l'ensemble de ses croyances permettent d'effectuer cette action, c'est-à-dire qu'elles sont compatibles avec les préconditions et les effets de l'action. Si ce n'est pas le cas, nous devons établir un diagnostic sur les croyances de l'opérateur permettant d'expliquer ce choix d'action. Il s'agit alors d'un problème de *consistency-based diagnosis* selon Reiter [19]. En effet le principe de ce diagnostic est de supposer que lorsqu'un système fonctionne normalement, le comportement du système est alors prévisible et tout comportement déviant de cette prédiction signifie qu'une erreur est apparue dans le système. Retrouver les conflits entre le comportement observé et les règles du système revient à trouver un diagnostic plausible. Dans notre cas, si un opérateur effectue une action en contradiction avec ses règles de raisonnement et croyances, alors retrouver les conflits revient à trouver les ignorances de l'opérateur. Or Wassermann [25] montre que ce problème de diagnostic peut être résolu à l'aide d'une révision de croyance dite « minimale ». Ainsi, le mécanisme de révision de croyances peut être utilisé à la fois pour prendre en compte les nouvelles observations et pour diagnostiquer les actions qui semblent « erronées » pour un observateur omniscient.

3.4 Révision de croyance

3.4.1 Principe général. La révision de croyance permet de retrouver la cohérence des croyances d'un agent face à une nouvelle information contradictoire [11]. Dans notre modèle, ce mécanisme est l'outil principal qui permet de transiter d'un état mental à un autre et donc de retrouver un raisonnement rationnel possible.

Néanmoins, toutes les révisions de croyance ne se valent pas. Considérons l'exemple suivant :

$$\begin{array}{lcl} R_1 & \equiv & (\text{pluie}_t \wedge \text{froid}_t) \Rightarrow \text{neige}_t \\ B_0 & \equiv & \{\text{froid}_0, \text{pluie}_0, R_1\} \\ Obs_1 & \equiv & \{\neg\, \text{neige}_0\} \end{array}$$

L'agent n'observe pas de neige alors qu'il croit qu'il pleut et qu'il fait froid (et donc, par application de R_1, qui est dans B_0, qu'il neige) : il y a donc une inconsistance. Tout sous-ensemble de $B_0 \cup Obs_1$ est une révision de croyance possible. Ainsi, on peut choisir d'ignorer l'observation, ou de renoncer à l'une de ses croyances sur le froid ou la pluie, ou d'abandonner la règle de déduction R_1. Mais on peut aussi choisir d'ignorer à la fois R_1 *et* une croyance : du point de vue de la logique mathématique, il n'y a pas de raison de choisir l'un plutôt que l'autre. Pourtant, du point de vue de la cognition humaine, ignorer l'ensemble des croyances pour prendre en compte la nouvelle information semble excessif.

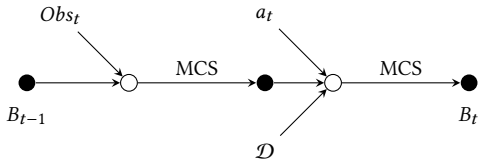
Ce problème a été étudié par Carlos Alchourron, Peter Gärdenfors et David Makinson [3] et a donné lieu à la théorie AGM (d'après le nom des chercheurs). Leur proposition est de définir un ensemble d'axiomes qui caractérisent une révision *minimale* des croyances. Dans notre cas, la solution qui consiste à ignorer $\{\text{pluie}_0, \text{froid}_0\}$ n'est pas minimale car il suffit d'éliminer pluie_0 ou froid_0. En d'autre termes, nous cherchons une correction minimale à apporter au système $B_0 \cup Obs_1$. On parle de *Minimal Correction Set* (MCS).

Plus formellement, pour un système $\Phi = \{\phi_1, \phi_2 \dots \phi_n\}$ donné, on peut dire que $M \subseteq \Phi$ est un MCS de Φ si et seulement si :

- $\Phi \setminus M$ est consistant
- $\forall \phi_i \in M, (\Phi \setminus M) \cup \{\phi_i\}$ est inconsistant

Dans notre exemple $\{\text{pluie}_0\}$ est un MCS, $\{\text{froid}_0\}$ en est un autre, mais $\{\text{pluie}_0, \text{froid}_0\}$ n'en est pas un. Nous utilisons l'algorithme de Liffiton [13] pour calculer l'ensemble des MCS d'un système. Cet algorithme présente l'avantage de définir un sous ensemble de Φ de croyances qu'il est impossible d'ignorer. Nous utilisons ceci pour interdire de réviser les actions effectuées par l'agent ainsi que l'ensemble des règles de cohérence C.

3.4.2 Application à notre modèle. Le calcul des états de croyance successifs de l'agent est illustré sur la figure suivante :



Nous ajoutons d'abord les observations. Si c'est inconsistant (point blanc), nous recherchons un MCS pour le rendre consistant (point noir). Ensuite, nous ajoutons les actions et les désirs instanciés pour l'instant suivant et nous calculons un nouvel MCS pour effectuer le *consistency-based diagnosis*.

À chaque étape de correction, nous n'obtenons pas nécessairement un unique MCS. Comme sur l'exemple météorologique précédent, par définition d'un MCS, il y a généralement plusieurs corrections possibles (ignorer la pluie, ignorer l'observation, ignorer $R_1 \dots$). Nous n'obtenons donc pas un schéma linéaire comme sur la figure ci-dessus, mais une arborescence d'états B_i où chaque embranchement correspond à un « choix de révision » possible de l'opérateur pour maintenir la cohérence. Chaque étage de l'arbre à un temps t de la simulation et chaque chemin de la racine vers une feuille définit un scénario comportemental (sur le plan cognitif). Notre objectif est de déterminer, parmi tous ces scénarios, ceux qui sont *plausibles*, c'est-à-dire qui correspondraient à un comportement possible d'un opérateur humain pour expliquer la décision d'action irrationnelle survenue lors de l'accident. Dans la prochaine section, nous montrons comment nous pouvons utiliser les biais cognitifs pour extraire les scénarios les plus plausibles.

3.4.3 Implémentation. Nous avons implémenté notre modèle sous le langage SMT-LIB [1] en utilisant le solveur z3 de Microsoft Research [2]. Cela nous permet d'avoir l'avantage des opérateurs incrémentaux "push" et "pop" pour interroger le solveur sans recharger le modèle à chaque fois. Nous chargeons les définitions des prédicats et des règles une fois, puis nous "pushons" les assertions correspondantes à chaque requête dans le cadre de notre algorithme de construction d'arbre MCS.

De plus nous avons définit un langage de modélisation avec une grammaire ANTLR (*ANother Tool for Language Recognition*), et nous avons écrit un programme pour traduire le modèle à partir de ce langage vers SMT-LIB. Cela nous permet d'écrire et d'analyser les modèles de manière systématique et plus efficace.

4 LES BIAIS COGNITIFS

4.1 Les biais cognitifs en sciences humaines

Trouver une explication par un raisonnement rationnel à un comportement irrationnel a été largement étudié en sciences humaines dans le domaine des biais cognitifs. En effet selon Tversky et Kahneman [23], l'humain utilise des raccourcis de pensée (heuristiques) pour compenser sa *rationalité limitée*, c'est à dire ses limitations en terme de mémoire et de temps. Ces heuristiques peuvent être très efficaces mais peuvent conduire à des erreurs. Nous parlons alors de biais cognitifs.

Dimara [9] 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 [16]. Par exemple, Takano et Reason [21] montrent que dans des accidents nucléaires au Japon et aux États-unis, l'excès de confiance, le biais de confirmation (*i.e.* la tendance à rechercher des informations en accord avec nos croyances et à rejeter les informations contradictoires [17]) et l'effet de récence (*i.e.* la facilité de se rappeler du dernier élément entendu) ont joué un rôle.

Les biais cognitifs peuvent donc conduire à des comportements irrationnels prévisibles. Notre proposition est que ces biais offrent une explication plausible à des accidents dans des processus de prise de décision. Nous présentons maintenant comment les intégrer dans notre modèle logique.

4.2 Les biais cognitifs dans notre modèle

Notre modèle considère que l'opérateur est omniscient d'un point de vue de la logique formelle : il connaît toute les conséquences logiques de sa base de croyance. Par exemple, si l'opérateur croit A et $A \Rightarrow B$ alors il croit B. Au vu de la littérature sur les biais cognitifs, cette hypothèse est bien trop forte pour représenter le raisonnement humain. Néanmoins, sous cette hypothèse d'omniscience, lorsqu'une action irrationnelle est effectuée (*i.e.* une action dont les préconditions ou les effets créent une inconsistance dans le système logique), nous calculons des MCS qui définissent des « ignorances possibles » sur la base de croyances permettant d'effectuer cette action tout en restant rationnel. Nous pouvons voir chaque MCS comme le retrait d'une part d'omniscience de l'opérateur menant à des états mentaux plausibles. C'est pourquoi nous postulons que ces ignorances peuvent traduire l'utilisation d'heuristiques, et par conséquent la présence de biais cognitifs dans la prise de décision de l'opérateur. En effet, les biais cognitifs étant des déviations systématiques du raisonnement vers un comportement irrationnel, la recherche des biais dans notre modèle logique revient à trouver quelles sont les propositions logiques ignorées pour que l'opérateur rationnel effectue une action irrationnelle.

Le choix de faire une correction *minimale* dans le système logique définit déjà un premier niveau de filtre sur l'ensemble des heuristiques de pensée. Il peut être interprété comme une minimisation du coût cognitif de la révision, bien que cette notion de coût cognitif soit probablement difficile à capturer à l'aide de la seule logique formelle. De plus, cette hypothèse pourrait être considérée comme trop forte car on peut imaginer qu'un être humain ignore pas seulement un ensemble minimale mais peut être un ensemble légèrement plus grand. Malgré cela, ce premier niveau permet de réduire considérablement l'ensemble des révisions possibles avant de les analyser sous l'angle des biais cognitifs. L'ensemble des scénarios d'états mentaux que nous obtenons en calculant les MCS sur toute la durée du scénario contient en effet beaucoup de scénarios, dont un certain nombre ne sont pas forcément plausibles du point de vue de la cognition humaine comme nous le montrerons dans la prochaine section à travers l'exemple du vol Rio-Paris. Notre proposition est de nous appuyer sur l'étude des biais cognitifs pour filtrer les scénarios les plus plausibles. Pour cela, nous recherchons des *motifs de MCS* qui peuvent correspondre à des biais en analysant de manière systématique la combinaison des croyances, observations, règles de raisonnement et MCS. Nous donnons ci-après, à l'aide de quelques exemples concrets, un ensemble non exhaustif de motifs de biais que nous pouvons retrouver à l'aide de notre modèle.

4.2.1 Biais d'attention. Le biais d'attention est défini comme une sélection des informations perçues en fonction de facteurs de préoccupation ou émotionnels [14]. Quand l'opérateur donne une préférence à une observation qui ne satisfait pas l'un de ses désirs, alors son attention est focalisée sur cette observation. De ce fait un MCS représentant un biais d'attention est défini par les conditions suivantes :

(1) $p, q \in Obs_t$ — p et q sont deux observations
(2) $\exists d \in \mathcal{D}.\ (B_{t-1} \wedge p) \Rightarrow \neg d$ — p ne satisfait pas un des désirs
(3) $q \in \mathrm{MCS}_t$ — q est ignorée

Par exemple :

$$
\begin{array}{lcl}
Obs_1 & \equiv & \{\mathrm{alarme}_1, \mathrm{reserve}_1\} \\
\mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R_1 & \equiv & \mathrm{alarme}_t \wedge \neg\, \mathrm{reserve}_t \Rightarrow \mathrm{panne}_t \\ R_2 & \equiv & \mathrm{alarme}_t \wedge \mathrm{reserve}_t \Rightarrow \neg\, \mathrm{panne}_t \\ R_3 & \equiv & [\mathrm{panne}_t]\, \mathrm{atterrirUrgence}_t \end{array} \right\} \\
\mathcal{D} & \equiv & \{\neg \mathit{panne}\} \\
a_1 & \equiv & \mathrm{atterrirUrgence}_1 \\
\mathrm{MCS}_1 & \equiv & \{\mathrm{reserve}_1\}
\end{array}
$$

Ici l'opérateur a le désir de ne pas tomber en panne d'essence. Il sait qu'il tombe en panne quand l'alarme sonne et qu'il n'a plus de réserve. Malgré des observations qui lui indiquent qu'il a une réserve et donc n'est pas en panne d'essence, il décide quand-même d'atterrir en urgence. Nous trouvons alors un MCS qui indique une attention focalisée sur l'alarme au point d'ignorer la réserve disponible.

1. http://smtlib.cs.uiowa.edu/
2. https://github.com/Z3Prover/

4.2.2 Biais d'engagement. L'escalade d'engagement est la tendance à persister dans une comportement irrationnel malgré des résultats de plus en plus négatifs [20]. Dans notre modèle nous considérons qu'un biais d'engagement est présent quand un opérateur rejette les effets d'une action précédente et continue d'appliquer cette action.

(1) $\exists d \in \mathcal{D}, B_{t-1} \Rightarrow \neg\, \mathrm{d}_{t-1}, B_t \Rightarrow \neg\, \mathrm{d}_t$ — Il existe un désir d non satisfait à $t-1$ et t.

(2) $\mathrm{a}_t = \mathrm{a}_{t-1}$ — même action choisie a à $t-1$ et t.

(3) $B_{t-1} \Rightarrow \mathrm{d}_t$ — Action a_{t-1} devait avoir satisfait d (n'oubliez pas que $\mathrm{a}_{t-1} \in B_{t-1}$)

(4) $B_t \Rightarrow \mathrm{d}_{t+1}$ — Action a_t est supposé satisfaire d

(5) $\mathrm{R}_k^a \in \mathrm{MCS}_t$ avec R_k^a la règle définissant l'effet de l'action a. L'opérateur a dû ignorer ces effets pour être en accord avec les observations de t (qui lui disent que a_{t-1} n'a pas fonctionné).

Considérons l'exemple suivant :

$$\begin{array}{lcl} Obs_1 & \equiv & \mathrm{vitesse}_1 \quad Obs_2 \equiv \mathrm{vitesse}_2 \\ \mathcal{R} & \equiv & \{\ R_1 \equiv \neg\, \mathrm{verglas}_t \Rightarrow (\mathrm{freiner}_t :: \neg\, \mathrm{vitesse}_{t+1})\ \} \\ \mathrm{a}_1 & \equiv & \mathrm{freiner}_1 \quad \mathrm{a}_2 \equiv \mathrm{freiner}_2 \\ \mathrm{MCS}_2 & \equiv & \{R_1\} \end{array}$$

Dans cet exemple un conducteur essaye de réduire sa vitesse en freinant, ce qui ne marche pas (probablement car il est en train de glisser sur du verglas). Malgré que sa décision n'est pas la bonne, il réessaye pourtant de freiner au pas de temps 2. Le MCS ignore l'effet du freinage, permettant à l'opérateur d'effectuer une nouvelle fois l'action malgré l'inconsistance entre l'effet attendu et la nouvelle information.

4.2.3 Excès de confiance. L'excès de confiance correspond à la tendance à considérer son jugement comme bien plus précis et performant qu'il ne l'est vraiment [15]. Dans notre modèle, Quand l'opérateur prédit un état du monde puis rejette toute information n'allant pas dans ce sens, il est dans un excès de confiance.

(1) $p \in Obs_t$ — p est une observation

(2) $(B_{t-1} \wedge p) \Rightarrow \bot$ — incohérente avec la prédiction

(3) $p \in \mathrm{MCS}_t$ — p est ignorée

Par exemple :

$$\begin{array}{lcl} Obs_1 & \equiv & \mathrm{nuages}_1 \quad Obs_2 \equiv \neg\, \mathrm{pluie}_2 \\ \mathcal{R} & \equiv & \{R_1 \equiv \mathrm{nuages}_t \Rightarrow \mathrm{pluie}_{t+1}\} \\ \mathrm{MCS}_2 & \equiv & \{\neg\, \mathrm{pluie}_2\} \end{array}$$

4.2.4 Biais de confirmation. Le biais de confirmation est la tendance à préférer les informations cohérentes avec nos croyances plutôt que celles qui jouent en défaveur de nos opinions [17]. Nous considérons que ce biais s'exprime dans notre modèle quand l'opérateur a le choix entre deux informations contradictoires et que l'information qui confirme les croyances est préférée. Nous entendons par confirmer les croyances, une information qui permet de déduire des croyances déjà connues par l'opérateur.

(1) $p, q \in Obs_t$ — p et q sont deux observations

(2) $\exists R_k \in B_{t-1}$ $\left\{ \begin{array}{rcl} (B_{t-1} \setminus R_k) \wedge p \wedge q & \nRightarrow & \bot \\ B_{t-1} \wedge p \wedge q & \Rightarrow & \bot \end{array} \right.$ — p et q sont contradictoires par la règle R_k

(3) $\exists B \subseteq B_{t-1}$ $\left\{ \begin{array}{rcl} (B_{t-1} \setminus B) & \nRightarrow & B \\ (B_{t-1} \setminus B) \wedge p & \nRightarrow & \bot \\ (B_{t-1} \setminus B) \wedge p & \Rightarrow & B \end{array} \right.$ — p confirme certaines croyances

(4) $(p \in \mathrm{MCS}_t \wedge q \in \mathrm{MCS}_t) \vee R_k \in \mathrm{MCS}_t$ — p est préférée à q, ou R_k est ignorée

Par exemple :

$$\begin{array}{lcl} Obs_1 & \equiv & \{\mathrm{voyantVert}_1\} \\ Obs_2 & \equiv & \{\neg\, \mathrm{voyantRouge}_2, \mathrm{alarme}_2\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lcl} R_1 & \equiv & \mathrm{voyantVert}_t \Rightarrow \neg\, \mathrm{panne}_t \\ R_2 & \equiv & \neg\, \mathrm{voyantRouge}_t \Rightarrow \neg\, \mathrm{panne}_t \\ R_3 & \equiv & \mathrm{alarme}_t \Rightarrow \mathrm{panne}_t \end{array} \right\} \\ \mathrm{MCS}_2 & \equiv & \{\mathrm{alarme}_2\} \vee \{R_3\} \end{array}$$

Ici l'observation du voyant vert permet de conclure qu'il n'y a pas de panne au temps 1. L'observation qu'il n'y a pas de voyant rouge au temps 2 permet la confirmation qu'il n'y a pas de panne. Si je retire le voyant vert de mes croyances je peux retrouver $\neg\, \mathrm{panne}_t$ par $\neg\, \mathrm{voyantRouge}_t$. Nous trouvons alors un MCS donnant plus de poids à l'information qui confirme en ignorant la contradiction avec cette croyance.

4.2.5 Conclusion. Les exemples ci-dessus montrent qu'il est possible de définir des motifs de biais dans les scénarios de raisonnement rationnel au sein du système croyances, observations, règles de raisonnement et MCS. Notre proposition est qu'un scénario composé de biais cognitif est une explication plus plausible pour un comportement irrationnel.

4.3 Application au cas du vol Rio-Paris

4.3.1 Explication de la situation. Pour illustrer notre méthodologie, nous modélisons une version simplifiée du crash du vol Rio-Paris en nous appuyant sur le rapport du BEA [1]. Quatre dispositifs principaux ont joué un rôle dans cet accident :

- Le *directeur de vol* (« Flight Director » ou « FD ») qui indique au pilote quelle manœuvre effectuer pour rejoindre la trajectoire programmée ;
- 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 2) ;
- L'altimètre qui donne une *vitesse verticale* (« Vertical Speed » ou « $\mathrm{V_z}$ »), 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).

De plus, les pilotes ont ressenti au début de l'incident des vibrations de l'appareil (« buffeting ») qu'ils ont mal interprétées.

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 2)). Au contraire, lorsque l'appareil est en sur-vitesse, il faut tirer sur le manche pour relever le nez de l'appareil [6]. C'est la confusion entre ces deux situations, en l'absence d'indications claires des instruments et sans visibilité externe, qui a conduit l'appareil à s'écraser.

4.3.2 Modélisation du problème. Nous pouvons résumer la situation de la manière suivante :

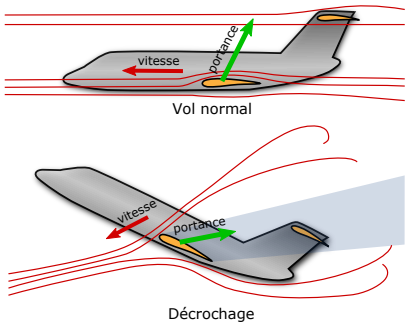


Figure 2: Décrochage (*source : Caliver, Wikimedia Commons*)

(t=1) Les indicateurs de vol affichent une accélération soudaine, l'alarme de décrochage s'active et une vibration 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.

Notre modèle permet de proposer des explications pour le comportement du pilote à l'aide des biais cognitifs et de la révision de croyances. Pour commencer, nous pouvons représenter les observations et les actions du pilote de la manière suivante :

$$\mathcal{B}elief_{init} \equiv \left\{ \begin{array}{l} \neg\,\text{alarme}_0, \neg\,\text{buffet}_0, \\ \neg\,\text{decrochage}_0, \neg\,\text{survitesse}_0 \end{array} \right\}$$

$$Obs_1 \equiv \{\text{buffet}_1, \text{alarme}_1, \text{acceleration}_1\}$$

$$Obs_2 \equiv \left\{ \begin{array}{l} \neg\,\text{buffet}_2, \neg\,\text{alarme}_2, \neg\,\text{acceleration}_2, \\ \text{FD}(\text{pull})_2, \text{V}_\text{z} \uparrow_2 \end{array} \right\}$$

$$Obs_3 \equiv \left\{ \begin{array}{l} \neg\,\text{buffet}_3, \neg\,\text{alarme}_3, \neg\,\text{acceleration}_3, \\ \text{FD}(\text{pull})_3, \text{V}_\text{z} \uparrow_3 \end{array} \right\}$$

$$Obs_4 \equiv \left\{ \begin{array}{l} \neg\,\text{buffet}_4, \text{alarme}_4, \neg\,\text{acceleration}_4, \\ \text{FD}(\text{pull})_4, \text{V}_\text{z} \uparrow_4 \end{array} \right\}$$

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

Pour comprendre ces observations et ces actions, il faut modéliser une partie de la connaissance des pilotes, ce qui correspond aux règles $\mathcal{R}$ suivantes :

$R_1 \equiv \text{buffet}_t \Rightarrow \text{decrochage}_t$
les vibrations indiquent un décrochage
$R_2 \equiv \text{alarme}_t \Rightarrow \text{decrochage}_t$
l'alarme de décrochage indique un décrochage
$R_3 \equiv \text{acceleration}_t \Rightarrow \text{survitesse}_t$
une accélération est un indicateur de sur-vitesse
$R_4 \equiv (\text{V}_\text{z} \uparrow_t \wedge \neg\,\text{decrochage}_t) \Rightarrow \text{survitesse}_t$
hors décrochage, l'augmentation de la V_z correspond à une survitesse
$R_5 \equiv \text{survitesse}_t \Rightarrow (\text{pull}_t :: \neg\,\text{survitesse}_{t+1})$
tirer le manche résout la sur-vitesse
$R_6 \equiv \text{decrochage}_t \Rightarrow (\text{push}_t :: \neg\,\text{decrochage}_{t+1})$
pousser le manche résout le décrochage
$R_7 \equiv \text{FD}(\text{pull})_t \Rightarrow \text{pull}_t$
l'opérateur devrait tirer sur le manche lorsque le directeur de vol le demande.

De plus, nous ajoutons deux désirs :

$$\mathcal{D} \equiv \{\neg\,\text{decrochage}, \neg\,\text{survitesse}\}$$

Enfin, pour que notre système logique soit complet, il faut lui indiquer que certaines choses sont incompatibles à l'aide des règles C (non-ignorables par le MCS) :

$C_1 \equiv (\text{decrochage}_t \wedge \text{survitesse}_t) \Rightarrow \bot$
décrochage et survitesse s'excluent l'un l'autre ;
$C_2 \equiv (\text{push}_t \wedge \text{pull}_t) \Rightarrow \bot$
idem pour tirer et pousser le manche.

4.3.3 Présentation et analyse des résultats. Le rapport du BEA met en avant les facteurs suivants pour expliquer le comportement du pilote :

(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)

En calculant l'ensemble des MCS sur les 4 pas de temps, notre modèle nous retourne 903 scénarios (ce qui correspond à un facteur de branchement moyen de 2.34 dans notre arbre de croyances). L'analyse des biais dans ces scénarios nous ont mené à identifier trois familles :

— Les scénarios dont les motifs de biais dans les MCS font ressortir l'ensemble des facteurs mis en évidence par le BEA ;
— Les scénarios dont les motifs de biais 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 MCS ne semblent pas traduire un comportement plausible

Scénario conforme à l'analyse du BEA. Dans la première famille nous trouvons par exemple le scénario suivant :

$$\begin{array}{lcl} MCS(t=1) & \rightarrow & R_1, Obs_1(\text{alarm}_1) \\ MCS(t=2) & \rightarrow & R_5 \\ MCS(t=3) & \rightarrow & Obs_3(\text{FD}(\text{pull})_3) \\ MCS(t=4) & \rightarrow & R_2 \end{array}$$

L'ignorance de la règle R1 (qui associe le buffet au décrochage) au temps 1 correspond au facteur (1). La sélectivité attentionnelle (2) fait référence au biais d'attention que nous avons présenté dans la section 4.2.1 et que nous retrouvons aussi au temps 1 : l'observation de l'alarme est ignorée et l'opérateur porte toute son attention sur la sur-vitesse par l'accélération de l'appareil.

Ce scénario d'état mental propose aussi l'ignorance de la règle R_5 au pas de temps 2, règle qui stipule que l'action « pull » devrait faire sortir l'appareil de la situation de sur-vitesse. L'action n'a pas eu cet effet au vu des observations et l'opérateur a dû ignorer cette règle. Le rapport de la BEA ne fait pas mention d'un tel raisonnement mais nous pouvons néanmoins faire un rapprochement avec le biais *d'engagement* [20] qui consiste à persister dans un même comportement avec des résultats de plus en plus négatifs.

Enfin l'ignorance au temps 3 de l'observation $\text{FD}(\text{pull})_3$ (c'est-à-dire le non-respect des consignes données par le directeur de vol) peut être interprété comme le fait que le pilote se rend compte que les directives de l'appareil sont mauvaises et change de stratégie. Mais au temps 4 l'alarme de décrochage se rallume et l'opérateur retient bien cette observation dans notre scénario (à t=4, le MCS ne contient pas l'observation de l'alarme). Pour maintenir la consistance du système, il ignore la règle R_2 qui fait le lien entre l'alarme et la situation de décrochage : cela correspond à notre motif de biais de confirmation présenté dans la section 4.2.4 et cela revient à ne pas considérer l'alarme comme pertinente, comme proposé par le BEA dans le facteur (3).

Les facteurs mis en évidence par le BEA sont donc bien retrouvés mais ce scénario met aussi en évidence que le *biais d'engagement* a pu jouer un rôle avec le biais d'attention et de confirmation.

Un autre scenario conforme à l'analyse du BEA. Toujours dans la première famille nous trouvons l'exemple suivant :

$$\begin{array}{lcl} MCS(t=1) & \rightarrow & R_1, Obs_1(\text{alarme}_1) \\ MCS(t=2) & \rightarrow & Obs_2(\text{V}_\text{z} \uparrow_2) \\ MCS(t=3) & \rightarrow & R_5, Obs_3(\text{FD}(\text{pull})_3) \\ MCS(t=4) & \rightarrow & R_2 \end{array}$$

Dans ce scénario nous retrouvons le biais d'attention au temps 1 et le biais de confirmation au temps 4. Néanmoins la règle R_5 n'est pas ignorée au temps 2, c'est l'observation de la vitesse verticale qui est ignorée. L'ignorance de cette observation traduit un excès de confiance de la part du pilote : il pense à $t = 1$ être en survitesse, et que son action pull lui permettra de sortir de cette situation. Ainsi le pilote ne fait pas attention à l'information incohérente avec sa prédiction, faisant confiance à son jugement. À $t = 3$, le pilote prend conscience de son erreur et considère que son action est mauvaise et que le directeur de vol donne de mauvaises indications. À $t = 4$ nous retrouvons un biais d'attention qui ne permet pas au pilote de considérer un décrochage. Nous retrouvons donc un scénario mêlant excès de confiance, biais d'attention et biais de confirmation qui permet d'expliquer le raisonnement du pilote et qui comporte les facteurs du BEA.

Scenario différents de l'analyse du BEA. Considérons le scénario suivant proposé par notre modèle comme explication possible du comportement du pilote :

$$\begin{array}{lcl} MCS(t=1) & \rightarrow & R_2, Obs_1(\text{buffet}_1) \\ MCS(t=2) & \rightarrow & R_5 \\ MCS(t=3) & \rightarrow & Obs_3(\text{FD}(\text{pull})_3) \\ MCS(t=4) & \rightarrow & \emptyset \end{array}$$

Dans ce scénario le biais d'attention porte sur la vibration et non sur l'alarme. De plus le pilote ignore la règle $R2$ et ne fait plus le lien entre l'alarme et le décrochage. Cette ignorance ne correspond pas à un biais de confirmation car aucune information à $t = 1$ ne confirme ou n'infirme les croyances. À première vue, l'ignorance d'une telle règle (en l'absence d'indices précédents) peut sembler forte mais cette hypothèse est citée par le BEA et pourrait s'expliquer par « la faible exposition [...] en formation continue (théorique et pratique) au phénomène de décrochage, à l'alarme STALL » (p.196). Si le pilote n'associe pas dans son esprit l'alarme au décrochage alors nous pouvons observer qu'il existe un scénario où cette non association suivi d'un biais d'engagement ne mène pas à un biais de confirmation. À $t = 4$, le pilote se retrouve dans une situation où il n'a aucune idée de ce qu'il doit faire et cette incompréhension du pilote est présente dans la transcription de l'exploitation de l'enregistreur phonique : « on a perdu le contrôle de l'avion on comprend rien on a tout tenté » (Annexe 1 p.28).

Un autre scénario de la même famille consiste à ignorer au temps 1 $\{R_6, Obs_1(\text{acceleration}_1)\}$: le pilote ignore la procédure à suivre lors d'un décrochage (peut être par manque de formation).

Scenarios devant être écartés. Notre moteur de MCS produit aussi certains scénarios absurdes, comme par exemple :

$$\begin{array}{lcl} MCS(t=1) & \rightarrow & R_1, R_2, R_3 \\ MCS(t=2) & \rightarrow & R_5 \\ MCS(t=3) & \rightarrow & R_7 \\ MCS(t=4) & \rightarrow & \emptyset \end{array}$$

Dans ce scénario, le pilote ignore au temps 1 toutes les règles lui permettant d'identifier un décrochage ou une sur-vitesse. Cela semble impossible pour un pilote professionnel : une action aléatoire pourrait avoir la même explication.

5 TRAVAUX CONNEXES

La littérature montre peu de travaux sur la représentation des biais dans un processus de prise de décision. En informatique décisionnelle, la plupart des travaux recherchent une solution optimale [22], et en simulation, les faits stylisés sont privilégiés [8]. Néanmoins nous pouvons citer Voison et al. [24] qui s'intéressent à la modélisation de l'impact des biais cognitifs dans les campagnes de vaccination. Ce modèle utilise un automate fini où chaque état représente le statut (infecté ou non) et l'opinion sur la vaccination. L'opinion est représentée par deux coûts : celui de la vaccination et celui de la transmission. Chaque transition entre états est calculée de manière biaisée, c'est-à-dire en fonction de l'opinion la plus présente dans la population (biais de conformisme) et des poids plus ou moins importants sur les coûts définissant l'opinion (biais de confirmation). Ce modèle a l'avantage d'être simple et permet de comprendre les mécanismes responsables des biais. Néanmoins le modèle n'est applicable que sur la situation de vaccination du fait que les croyances ne se résument qu'à une opinion.

Kulick et al.[12] cherchent à prédire le comportement d'un agent face à une opération stratégique (militaire, diplomatique, ...). Ils considèrent l'agent comme rationnellement limité et donc biaisé. En revanche les auteurs cherchent à prédire l'effet d'un raisonnement biaisé et non la dynamique qui a poussé l'agent à ce raisonnement. C'est pourquoi ils utilisent un *synthetic cognitive model* (SCM), un modèle en boite noire où des facteurs en entrée mènent à des prédictions de comportements. Notre modèle au contraire cherche à expliquer le raisonnement d'un comportement passé.

Arnaud et al. [4] se basent sur un modèle BDI pour représenter les biais cognitifs dans la situation de feux de forêt. Chaque biais cognitif est représenté par une fonction qui diminue ou augmente la probabilité d'une croyance en fonction d'une nouvelle information disponible pour l'agent. Ainsi pour prendre en compte un nouveau biais, il suffit d'écrire sa fonction probabiliste correspondante. Néanmoins comme vu subsection 4.1, 151 biais sont répertoriés ce qui donne 151 fonctions possibles sans avoir l'assurance que celles-ci ne se recouvrent pas car il n'existe pas, à ce jour, de consensus sur la taxonomie des biais [5]. C'est pourquoi, à l'inverse de Arnaud et al., nous basons notre modèle sur une approche de diagnostic afin de capturer le plus de biais possibles et de ne pas limiter l'explication d'un raisonnement à un seul biais.

Enfin Dutilh Novaes et al. [18], dans le domaine de la philosophie et de la logique, montrent empiriquement qu'un modèle basé sur la révision de croyances minimale (respectant AGM) permet de prédire le comportement d'un agent sur des tâches liées au *Belief bias* (la tendance à juger des arguments en fonction de la plausibilité de la conclusion et non à quel point ils supportent la conclusion). Notre modèle, se basant aussi sur la révision de croyances minimale, se rapproche de ces travaux. Néanmoins les auteurs proposent un modèle prédictif pour un seul et unique biais là où nous proposons un modèle explicatif de plusieurs biais.

6 CONCLUSION

Notre modèle utilise le diagnostic basé sur la révision de croyances et les biais cognitifs pour identifier un ensemble de scénarios d'états mentaux rationnels permettant d'expliquer un comportement irrationnel. Il propose pour chaque scénario la trace des évènements et les choix de révision de l'agent correspondants aux différents biais. De plus ce modèle a l'avantage de reposer sur une base théorique forte nourrie de plus de 35 ans de recherche [10] facilitant l'ajout d'extensions.

Néanmoins nous pouvons mettre en évidence plusieurs limites à ce modèle qui seront sujet à de futurs travaux. Dans un premier temps, nous souhaitons étendre notre ensemble de motifs de biais et implémenter un algorithme de recherche systématique qui permettra de faire une présentation claire des biais retenus. En effet, sur l'exemple du vol Rio-Paris, l'analyse est faite à la main sur la base des quelques motifs que nous voulions retrouver à partir de l'analyse du BEA. Notre objectif est ainsi de concevoir une taxonomie de biais en nous inspirant des taxonomies présentes dans la littérature en sciences humaines, tout en développant notre classification des biais propre à notre modèle, et ce de manière à ce que nos motifs forment une partition des MCS possibles.

Une deuxième perspective à notre travail, dans l'optique de proposer un outil pour des non-spécialistes, serait de classer les scénarios par ordre de probabilité ou de plausibilité. Nous voudrions pour cela utiliser les facteurs situationnels et de personnalité associés aux biais, qui poussent les individus à utiliser une heuristique plutôt qu'une autre.

Enfin, nous voudrions étendre notre modèle pour y intégrer les émotions. Plusieurs travaux (voir par exemple [14] sur le biais d'attention), font état de l'importance des émotions dans les biais. Un pilote ayant une expérience traumatisante d'une panne d'essence fera tout pour éviter de retomber dans le même contexte (et donc possiblement prendra d'autres risques). Nous souhaitons nous appuyer sur les modèles affectifs utilisant la logique BDI, comme [2, 7]. Notre modèle constitue en effet une base solide pour prendre en compte un large ensemble de facteurs d'irrationalité dans des situations de prise de décision tout en répondant aux limites des modèles des biais cognitifs actuels.

RÉFÉRENCES

[1] 2012. *BEA f-cp090601.* Technical Report. Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile.

[2] Carole Adam, Andreas Herzig, and Dominique Longin. 2009. A logical formalization of the OCC theory of emotions. *Synthese* 168, 2 (2009), 201–248.

[3] Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change : Partial meet contraction and revision functions. *The journal of symbolic logic* 50, 2 (1985), 510–530.

[4] Maël Arnaud, Carole Adam, and Julie Dugdale. 2017. The role of cognitive biases in reactions to bushfires. In *ISCRAM*. Albi, France.

[5] Andrea Ceschi, Arianna Costantini, Riccardo Sartori, Joshua Weller, and Annamaria Di Fabio. 2019. Dimensions of decision-making : An evidence-based classification of heuristics and biases. *Personality and Individual Differences* 146 (2019), 188–200.

[6] Stéphane Conversy 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*. ACM, 60–69.

[7] Mehdi Dastani and Emiliano Lorini. 2012. A logic of emotions : from appraisal to coping.. In *AAMAS*. 1133–1140.

[8] Paul Davidsson. 2002. Agent based social simulation : A computer science view. *Journal of artificial societies and social simulation* 5, 1 (2002).

[9] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic. 2020. A Task-Based Taxonomy of Cognitive Biases for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 26, 2 (2020), 1413–1432.

[10] Eduardo Fermé and Sven Ove Hansson. 2011. AGM 25 Years : Twenty-Five Years of Research in Belief Change. *Journal of Philosophical Logic* 40 (04 2011), 295–331.

[11] Peter Gärdenfors and Hans Rott. 1995. *Belief Revision.* Vol. 4. 35–132.

[12] Jonathan Kulick and Paul K Davis. 2003. Modeling Adversaries and Related Cognitive Biases. *Modeling Adversaries and Related Cognitive Biases* (2003).

[13] Mark H Liffiton and Karem A Sakallah. 2008. Algorithms for computing minimal unsatisfiable subsets of constraints. *Journal of Automated Reasoning* 40, 1 (2008), 1–33.

[14] Colin MacLeod, Andrew Mathews, and Philip Tata. 1986. Attentional bias in emotional disorders. *Journal of abnormal psychology* 95, 1 (1986), 15.

[15] Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological review* 115, 2 (2008), 502.

[16] Atsuo Murata, Tomoko Nakamura, and Waldemar Karwowski. 2015. Influence of cognitive biases in distorting decision making and leading to critical unfavorable incidents. *Safety* 1, 1 (2015), 44–58.

[17] Raymond S Nickerson. 1998. Confirmation bias : A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[18] Catarina Novaes and Herman Veluwenkamp. 2016. Reasoning Biases, Non-Monotonic Logics and Belief Revision. *Theoria* 83 (12 2016).

[19] Raymond Reiter. 1987. A theory of diagnosis from first principles. *Artificial Intelligence* 32, 1 (1987), 57 – 95.

[20] Barry M. Staw. 1996. *The escalation of commitment : An update and appraisal.* Cambridge University Press, 191–215.

[21] Kenichi Takano and James Reason. 1999. Psychological biases affecting human cognitive performance in dynamic operational environments. *Journal of Nuclear Science and Technology* 36, 11 (1999), 1041–1051.

[22] Alexis Tsoukiàs. 2008. From decision theory to decision aiding methodology. *European Journal of Operational Research* 187, 1 (2008), 138–161.

[23] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty : Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.

[24] Marina Voinson, Sylvain Billiard, and Alexandra Alvergne. 2015. Beyond rational decision-making : modelling the influence of cognitive biases on the dynamics of vaccination coverage. *PloS one* 10, 11 (2015).

[25] Renata Wassermann. 2000. *An Algorithm for Belief Revision.* Technical Report.