Diagnostiquer les biais cognitifs

Valentin Fouillard, Safouan Taha, Nicolas Sabouret, Frédéric Boulanger

Diagnostiquer les biais cognitifs

Valentin Fouillard[1] Safouan Taha[2] Nicolas Sabouret[1] Frédéric Boulanger[2]

[1] Université Paris-Saclay, CNRS, LIMSI, 91400 Orsay France
[2] Université Paris-Saclay, CNRS, CentraleSupélec, LRI, 91405, Orsay, France

prenom.nom@universite-paris-saclay.fr

Résumé

De nombreux exemples d'accidents impliquant une prise de décision par un opérateur humain, montrent une incohérence entre les actions observées et les actions attendues. Dans cet article, nous proposons un modèle capable de fournir une explication de ce comportement à travers les biais cognitifs en se basant sur les propriétés de la révision des croyances AGM. Nous montrons qu'un opérateur respectant ces propriétés suffit à générer des comportements similaires aux biais. De plus nous montrons que caractériser un biais revient à trouver une fonction de sélection parmi les comportements générés.

Mots Clef

Biais cognitif, AGM, révision de croyance, diagnostic, prise de décision

Abstract

Many examples of accidents involving decision making by a human operator show an inconsistency between the expected actions and the observed actions. In this paper, we present a model that can explain this kind of behavior through cognitive biases, based on AGM belief revision properties. We show that an operator that behaves according to these properties can generate biases-like behaviors. Moreover we show that characterizing a bias is similar to finding a selection function among the generated behaviors.

Keywords

Cognitive biases, AGM, belief revision, diagnostic, decision making

1 Introduction

Le 1er juin 2009, le vol 447 d'Air France entre Rio et Paris s'écrase dans l'océan atlantique. C'est l'accident le plus meurtrier d'Air France avec 228 morts. Trois ans plus tard, en juillet 2012, paraît le rapport du Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile [10]. L'analyse montre le déroulé suivant pour l'accident :

— incohérence temporaire entre les vitesses mesurées, vraisemblablement à la suite de l'obstruction des sondes Pitot par des cristaux de glace, ayant entraîné notamment la déconnexion du pilote automatique ;
— situation de décrochage suite à la déconnexion du pilotage automatique ;
— maintien de la situation de décrochage par les pilotes par leurs actions.

De plus il peut être noté que l'alarme de décrochage a retenti plus de 75 fois sans que les pilotes identifient leur situation comme un décrochage. D'un point de vue extérieur et avec toutes les informations à notre disposition, ce comportement semble irrationnel. Pourtant si nous nous posons du point de vue du pilote avec les informations à sa disposition, ses actions semblent faire sens. En effet, le rapport montre que :

— l'alarme de décrochage sonnait de manière furtive et pouvait être considérée comme non pertinente ;
— il n'y avait pas d'information visuelle permettant de confirmer l'approche du décrochage après la perte des vitesses caractéristiques ;
— il y avait confusion possible avec une situation de survitesse ;
— des indications des directeurs de vol pouvaient conforter l'équipage dans ses actions, bien qu'elles soient inappropriées.

Nous voyons ici que nous avons à faire à une erreur de jugement de la situation qui diffère de la réalité et qui est confortée par les informations à la disposition des pilotes. Ce phénomène est largement étudié en psychologie dans le domaine des **biais cognitifs** et est présent jusque dans la méthode scientifique [23], la justice [16] ou en médecine [25]. Ainsi, comme le déclare Murata [22], si les ingénieurs oublient de prendre en compte et de comprendre les limitations humaines dans la conception des technologies modernes, la distorsion des décisions se produit inévitablement, aggravant les erreurs humaines et conduisant finalement à des incidents, des accidents, des collisions ou des catastrophes.

En informatique, la branche qui s'intéresse à modéliser et comprendre les erreurs dans les systèmes est le diagnostic [8]. Dans cet article, nous nous intéressons à un cas particulier de diagnostic. Il s'agit, à partir d'un comportement

attendu et d'un comportement observé (suite d'actions) qui semble irrationnel, d'apporter une explication en s'appuyant sur les biais cognitifs. Nous proposons pour cela de nous appuyer sur les modèles de révision de croyances pour déterminer quels biais peuvent être mis en œuvre pour mener à une erreur.
Dans la prochaine section, nous présentons la définition des biais cognitifs en sciences humaines et leur modélisation en informatique. Nous faisons le lien avec la rationalité telle qu'elle est représentée dans les modèles de révision de croyance. Dans la troisième section, nous montrons comment extraire des biais cognitifs à partir d'un ensemble de connaissances ayant conduit à une situation d'erreur. Nous illustrons ce fonctionnement sur un cas d'école. Nous montrons enfin dans les perspectives quelles sont les propriétés que devrait respecter une fonction de sélection de biais dans un tel modèle.

2 Travaux connexes

Dans cette première section nous définissons les biais cognitifs selon les sciences humaines et discutons des modèles informatiques de ce phénomène irrationnel. Nous détaillons ensuite la raison pour laquelle la révision de croyance est le mécanisme central de la rationalité, et son lien avec le diagnostic.

2.1 Les biais cognitifs

Selon Tversky et Kahneman [33], l'être humain utilise des raccourcis de pensée (heuristiques) afin de compenser ses limitations en mémoire et en temps. Ces heuristiques peuvent parfois mener à des erreurs, c'est-à-dire des comportements qui diffèrent de la théorie du choix rationnel. Les auteurs montrent par exemple que les individus fondent leur jugement sur des informations personnelles plutôt que statistiques. Todd [31] montre l'importance de ces heuristiques comme par exemple l'heuristique *Take the Best*. Dans le problème de choisir une option parmi un ensemble de possibilités selon un critère de comparaison, *Take the Best* revient à prendre la première information qui départage les possibilités avec ces informations rangées par ordre de validité. La validité est la probabilité conditionnelle qu'un objet tombe dans une catégorie selon l'information choisie. Todd montre que *Take the best* est plus performant en généralisation et utilise moins d'information pour départager qu'une régression multiple. Mais dans certains cas, la solution obtenue grâce à cette heuristique est sous-optimale : nous parlons alors de *biais cognitif*.
Benson [4] répertorie l'ensemble des biais cognitifs (175 biais) en 4 catégories reflétant les problèmes pour lesquels ces heuristiques apportent une solution :

- trop d'information,
- pas assez de sens,
- manque de temps,
- compromis entre l'oubli et le souvenir.

L'ensemble de ces catégories contient 20 sous-ensembles reflétant la stratégie utilisée pour répondre à ce problème.
Plusieurs études ont montré le rôle important de certains de ces biais dans la prise de décisions à forts risques. Par exemple, dans le domaine de l'aviation, Walmsley et Gilbey [35] effectuent plusieurs expérimentations dans lesquelles un pilote est informé de la météo prévue en amont, puis un changement du temps apparaît pendant son trajet. Les différentes expériences montrent que :

- le bulletin météo donné avant l'envol influence l'interprétation des indices météorologiques par les pilotes (*le biais d'ancrage*) ;
- il n'existe pas de preuve montrant que les pilotes favorisent une stratégie de recherche d'éléments contradictoires par rapport à une stratégie de recherche de confirmation (*biais de confirmation*) ;
- les pilotes évaluent positivement le fait d'avoir traversé une zone de conditions météorologiques instables si le résultat est positif (*outcome bias*).

Le biais d'ancrage est le fait de s'appuyer sur l'information reçue en premier pour prendre une décision. *Le biais de confirmation* consiste à privilégier ses idées préconçues. Enfin l'*outcome bias* consiste à évaluer trop vite une décision quand son effet est connu. De plus, Gibley et Hill [15] effectuent des expérimentations similaires mais sur le scénario d'un pilote perdu qui doit retrouver son chemin. Une majorité des pilotes utilisent une stratégie de confirmation et prennent la mauvaise direction, mettant en avant ici encore le biais de confirmation. Enfin, nous pouvons citer Malmquist [20] qui met en relation les biais cognitifs et des exemples célèbres d'accident d'aviation (le vol Air France 447 par exemple), ainsi que Brafman & Brafman [7] qui montrent que *l'aversion à la perte* a largement contribué au crash du vol KLM 4805, l'aversion à la perte consistant à attacher plus d'importance à une perte qu'à un gain de même montant.

Dans l'industrie nucléaire, Takano et Reason [30] se sont basés sur des analyses des facteurs humains concernant les accidents dans des centrales nucléaires aux États-Unis, au Japon et sur un simulateur. Leur étude montre le rôle de l'excès de confiance, l'effet de récence qui est la facilité de se rappeler du dernier élément entendu et le biais de confirmation dans l'évaluation d'une situation inattendue.

Enfin, Murata [22] dans le domaine de l'accidentologie, met en évidence l'importance de certains biais cognitifs, notamment l'excès de confiance, le biais d'optimisme qui consiste à croire que les évènements négatifs nous touchent moins que les autres, et le framing effect qui est le fait de choisir une option selon le fait quelle soit présenté positivement ou négativement

L'ensemble de ces études permet d'identifier des biais fréquents dans la prise de décision : biais de confirmation, effet de récence, aversion à la perte, excès de confiance, *etc*. Nous nous concentrons sur ces biais pour construire des diagnostics sur les comportements erronés des opérateurs humains.

2.2 Modèles informatiques des biais

La littérature en informatique montre peu de travaux portant sur la modélisation des biais dans la prise de décision. La plupart des travaux en informatique décisionnelle visent à calculer une solution optimale [32] tandis que les travaux en simulation cherchent souvent à retrouver des faits stylisés [9].

Cependant, nous pouvons citer Voison [34] qui s'intéresse à la modélisation de l'impact des biais cognitifs dans les campagnes de vaccination. Ce modèle est basé sur un modèle de croyance et deux fonctions distinctes représentant chacune des biais. Le modèle mathématique est simple mais ne modélise que deux biais cognitifs. Nous visons ici à avoir une approche plus générale et à proposer un modèle prenant en compte plusieurs biais.

Le travail de Kulick [18] cherche à modéliser l'effet d'une opération militaire sur une cible. Ce modèle se base sur un modèle de « boîte noire » avec des facteurs en entrée pour aboutir à des probabilités de comportements possibles. Néanmoins, celui-ci n'offre pas la possibilité d'intégrer des mécanismes de biais connus, et du fait de son fonctionnement, le modèle ne donne pas le mécanisme mental qui mène au comportement en sortie.

Enfin nous pouvons citer Arnaud et Adam [3] qui cherchent à modéliser les biais cognitifs lors des feux de forêts en se basant sur le modèle BDI [28]. Leur approche s'appuie sur un mécanisme de mise à jour de croyance en fonction de probabilités. Cette approche à l'avantage de pouvoir intégrer facilement de nouveaux biais dans le modèle et d'avoir une description claire du mécanisme mental effectué. Cependant, le choix d'une révision probabiliste des croyances de l'agent n'est pas conforme aux modèles de révision de croyances qui ont été largement étudiés et formalisés en logique.

Nous pensons qu'en nous appuyant sur un mécanisme de révision de croyances et un modèle de type BDI, nous pourrons bénéficier d'une base théorique forte tout en offrant une facilité d'intégration à des modèles existants.

2.3 Révision de croyance

La révision des croyances est le mécanisme qui permet de passer d'un état de croyance à un autre tout en gardant la rationalité de l'agent. Ce dernier a été largement étudié en informatique [14].

Pour pouvoir étudier les biais cognitifs, nous proposons de nous appuyer sur ces modèles pour construire des transitions rationnelles entre des états pouvant expliquer un comportement irrationnel. Il faut donc que chaque étape soit conforme au modèle de révision de croyance, tout en menant à une situation d'erreur de décision (par exemple, le maintien de l'appareil en situation de décrochage malgré les alarmes lors du crash Rio-Paris).

Le problème de la révision de croyance est de choisir les modifications à effectuer dans une base de connaissances pour prendre en compte une nouvelle information. Pour illustrer ce problème, nous prenons l'exemple proposé par *Gärdenfors et al.* dans leur livre fondateur [14]. Considérons la base de croyances suivante :

(1) L'oiseau attrapé dans la cage est un cygne (A)
(2) L'oiseau attrapé dans la cage vient de Suède (B')
(3) La Suède fait partie de l'Europe ($B' \rightarrow B$)
(4) Les cygnes européens sont blancs ($A \wedge B \rightarrow C$)

Nous pouvons en déduire alors que l'oiseau attrapé dans la cage est blanc (C). Imaginons maintenant que nous observons que l'oiseau est noir ($\neg C$). Nous devons alors réviser nos croyances afin de garder une cohérence dans notre base. Pour cela, nous devons éliminer une des croyances (1-4) mais d'un point de vue logique aucune préférence n'est donnée pour l'élimination : l'ensemble de la combinaison des quatre croyances est acceptable pour réviser notre croyance.

Bien que d'un point de vue logique l'élimination de toutes les anciennes croyances pour prendre en compte la nouvelle information soit recevable, elle n'est pas rationnelle. C'est pourquoi Alchourrón, Gädenfors et Makinson (AGM) [1] définissent des postulats veillant à la rationalité de la révision des croyances. Le but de ces propriétés est de retenir autant que possible nos anciennes croyances, en d'autres termes, nous ne voulons changer que le nécessaire : c'est le *changement minimal*.

Le modèle AGM définit trois opérations possibles pour le changement de croyances : la contraction ($\div$), l'expansion ($+$) et la révision ($*$). La contraction est le fait de retirer une croyance de la base. L'expansion consiste à ajouter une nouvelle croyance à la base et la révision est l'ajout d'une nouvelle connaissance dans la base mais tout en gardant sa cohérence. Ainsi la révision peut être construite à partir de la contraction et de l'expansion selon l'*identité de Levi* [1] :

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

où K est la base de croyance et p une nouvelle information.

La **contraction** est donc au cœur de cette question de révision de croyances et le modèle AGM caractérise ces fonctions de contractions rationnelles, appelées *partial meet contraction*. Ces travaux ont donné lieu à de nombreuses recherches [12] ainsi que plusieurs définitions de la révisions de croyances, dont il a été montré qu'elles sont équivalentes. En particulier, nous pouvons citer le *kernel contraction* [17] dont le principe est de calculer les ensembles minimaux impliquant la proposition à retirer de la base, et le *Epistemic Entrenchement* [13] qui définit une relation d'ordre sur les croyances de la base K pour abandonner en priorité les croyances avec un enracinement plus faible.

Dans notre article, nous utiliserons la méthode des Minimal Correction Set (MCS) qui consiste à calculer l'ensemble des corrections minimales à apporter à un ensemble de croyances insatisfiable pour le rendre satisfiable. Besnard [5] montre que MCS est le complément de *Maximal Subset Satisfiable*, ce qui revient a faire une *partial meet contraction* qui respecte donc AGM. Plus formellement :

$M \subseteq K$ est un MCS de K ssi :

— $K \backslash M$ est SAT
— $\forall p \in M, (K \backslash M) \cup \{p\}$ est UNSAT

Il existe dans la littérature de nombreux algorithmes capables de calculer un tel ensemble [21].

Lien entre le diagnostic et la révision des croyances. Wassermann [36] montre qu'un problème de diagnostic peut être traduit en un problème de révision de croyance. En effet, l'approche *consistency-based diagnosis* proposée par [29] consiste à décrire un système en termes de composants et des interactions entre ces derniers indépendamment de la tâche de diagnostic. Le but est alors de trouver le composant responsable lorsqu'un comportement anormal est détecté. L'approche consistency-based diagnosis est un système de trois ensembles logiques :

— SD qui est la description du système ;
— ASS qui sont les hypothèses ;
— OBS qui sont les observations.

L'ensemble $\Delta \subset ASS$ est un diagnostic pour le triplet (SD, ASS, OBS) ssi Δ est un ensemble minimal tel que $SD \cup OBS \cup (ASS \backslash \Delta)$ est consistant. Ainsi, si nous partons de l'hypothèse que tous les composants sont fonctionnels alors le triplet sera inconsistant et Δ revient à trouver et retirer le composant défectueux. Wassermann montre que nous pouvons utiliser AGM pour calculer Δ car cela revient à trouver une partial meet contraction, c'est à dire trouver l'ensemble maximal satisfaisable.

Un tel diagnostic est donc l'ensemble minimal de propositions à retirer pour retrouver la consistance. Notre problème de diagnostic d'accident consiste, à partir d'une situation donnée et d'une prise de décision irrationnelle, à déterminer quelle croyance a été ignorée (retirée) par l'agent à cause d'un biais cognitif. Nous pouvons donc nous reposer sur l'opérateur de révision de croyance pour déterminer les croyances inconsistantes.

Dans la suite de cet article, nous montrons comment cette approche de diagnostic à partir de révision de croyances peut être mise en œuvre pour extraire des biais cognitifs et expliquer des situations d'accident.

3 Contribution

Notre modèle part d'une situation où l'agent se comporte de manière irrationnelle. Nous avons comme supposition ses croyances et l'ensemble des observations que l'agent a pu faire. Nous cherchons une explication possible de son comportement. De ce fait, nous traduisons ce problème de diagnostic en un problème de révision de croyances en se basant sur AGM. Une fonction de révision de croyance AGM est divisée en deux parties :

(1) génération des solutions conformes à AGM ;
(2) sélection de la meilleure solution.

Nous allons montrer dans cet article que (1) permet d'avoir un ensemble comportant des biais cognitifs possibles (tout en préservant la rationalité de l'agent) et que caractériser un biais spécifique revient à sélectionner (2) la solution représentant ce biais. En d'autres termes, nous montrons ici que la révision de croyances est suffisante pour mettre en évidence des biais cognitifs. De plus, nous discuterons des caractéristiques nécessaires des fonctions de sélection de biais.

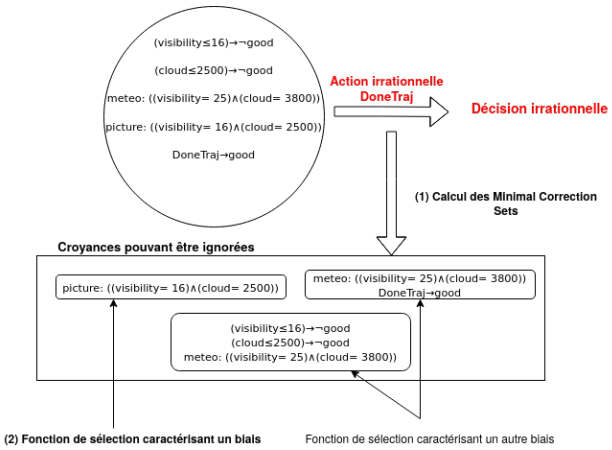

FIGURE 1 – Modèle du diagnostic des biais cognitifs

3.1 Exemple concret

Nous allons considérer le problème suivant, inspiré de l'expérimentation de Walmsley et Gibley [35] présentée dans la section précédente. À l'approche de son point d'arrivée, un pilote d'avion doit prendre en compte de nouvelles informations météorologiques moins favorables. Nous pouvons modéliser ce problème avec la base de croyances suivante :

(1) $(visibility \leq 16) \rightarrow \neg good$
Si la visibilité est en dessous de 16 km alors les conditions de vol sont mauvaises

(2) $(cloud \leq 2500) \rightarrow \neg good$
Si les nuages sont en dessous de 2500 pieds alors les conditions de vol sont mauvaises

(3) $meteo : ((visibility = 25) \wedge (cloud = 3800))$
La météo fournie au début du vol déclare une visibilité de 25km et des nuages à 3800 pieds

(4) $picture : ((visibility = 16) \wedge (cloud = 2500))$
La photographie reçue à l'approche du point d'arrivée montre une visibilité de 16 km et des nuages à 2500 pieds

(5) $DoneTraj \rightarrow good$
Je ne peux faire le trajet que si les conditions sont bonnes

Nous nous plaçons dans la situation où le pilote effectue le trajet (décision erronée) : $DoneTraj$ est nécessairement vrai. Le système [1-5] est d'entrée de jeu insatisfaisable car l'agent ne peut pas accepter plusieurs valeurs différentes pour la visibilité et les nuages (propositions 3 et 4). De plus si l'agent prend en compte les nouvelles informations qui lui ont été communiquées sur la photographie, le système reste insatisfaisable, du fait que $DoneTraj$ doit être vrai. Nous allons donc utiliser un algorithme de MCS pour calculer les corrections rationnelles possibles pour rendre cet ensemble satisfaisable. Chaque correction

correspond à une révision de croyance possible pouvant expliquer $DoneTraj$. Cela correspond à l'ensemble des croyances que l'agent a dû ignorer pour effectuer cette action. Dans notre exemple, nous nous retrouvons avec l'ensemble de solutions suivantes :

$$\{\{3,1,2\},\{3,5\},\{4\}\}$$

3.2 Extraction de biais cognitifs

Dans cette partie, nous allons discuter de chacune de ces solutions :

- $\{4\}$ représente le fait d'ignorer (involontairement ou volontairement) les informations météorologiques contenues dans la photo.
- $\{3,1,2\}$ représente le fait d'ignorer les connaissances sur le danger et d'ignorer (involontairement ou volontairement) les informations données par la météo au début du vol.
- $\{3,5\}$ représente le fait de considérer le danger comme négligeable pour faire le trajet. L'agent ignore ici encore les informations météorologiques initiales (pour garder une seule valeur possible pour les nuages et la visibilité).

Si nous gardons les hypothèses de l'étude où les informations sur les photos et les données météorologiques ne pouvaient être ignorées involontairement alors le cas $\{4\}$ vient capturer le comportement d'un biais d'ancrage mis en évidence dans l'étude. En effet, ignorer ces informations revient à placer trop d'importance dans la véracité des données météorologiques et oblige à ignorer les informations de la photo : "Specifically, pilots appeared to place too much importance on reports that may no longer be valid and then fail to adjust their perceptions in the face of evidence to the contrary (i.e. the presented images)" [35].

Ignorer la météo (3), c'est à dire donner une priorité de croyance sur la photographie, revient à faire deux choix de révision possibles. Soit l'agent ignore ses croyances sur les conditions de danger ($\{3,1,2\}$), soit il ignore le fait qu'il est nécessaire pour ce trajet que les conditions soient bonnes($\{3,5\}$). Bien que ces comportements ne ressortent pas dans l'étude de Walmsley et Gilbey, ils n'en restent pas moins intéressants. Par exemple ignorer $\{3,5\}$ peut être vu comme un excès de confiance : l'agent ne remet en cause ni les photographies ni le danger, mais considère que, dans son cas, les conditions pour faire le trajet sont suffisantes. Enfin $\{3,1,2\}$ peut correspondre à une aversion à la perte, c'est-à-dire que le fait de ne pas continuer le trajet est une perte trop importante aux yeux du pilote et qu'il est préférable d'ignorer le danger et de continuer. Néanmoins, il faut garder en mémoire que les solutions du MCS ne représentent pas uniquement des ignorances volontaires : il peut aussi s'agir d'oublis involontaires et donc de choix non-intentionnels de la part de l'agent.
Enfin, nous avons considéré dans toutes les hypothèses mentionnées ci-avant que l'agent percevait toutes les informations (ce qui était le cas dans l'étude de Walmsley et Gilbey).

3.3 Sélection d'une solution

Face à un comportement irrationnel, nous pouvons diagnostiquer par la révision de croyances des comportements équivalents au fonctionnement de certains biais cognitif. Ainsi, la révision de croyances par AGM suffit à retrouver des biais cognitifs. Néanmoins, si nous voulons définir quelle solution, parmi l'ensemble des possibilités, correspond à un biais en particulier, nous devons prendre en compte des caractéristiques qui ne sont pas présentes dans la modélisation de l'exemple. Par exemple, quels facteurs ont influencé le pilote pour prendre la décision d'ignorer les informations de la photographie plutôt que de prendre la décision d'ignorer la météo et le danger ? En d'autres termes, pourquoi la météo est-elle difficile à abandonner dans le premier cas et les informations dans la photographie dans le deuxième cas ? Ces points seront adressés dans de futurs travaux dont nous donnons les premières pistes dans la section suivante.

4 Travaux Futurs

Nous nous retrouvons avec un ensemble de solutions de révision dont le comportement de sortie peut correspondre à des biais. Néanmoins, nous n'avons pas défini une fonction de sélection permettant de dire quelle croyance ignorée correspond à un biais. Si nous nous référons aux résultats de l'étude [35], une majorité de pilotes a choisi d'ignorer $\{4\}$ et non $\{3,1,2\}$ ou $\{3,5\}$. De ce fait, il faut que la fonction de sélection soit une fonction de préférence sur l'ensemble des solutions. Or l'*epistemic entrenchement* [13] a ce fonctionnement comme vu à la section 2.3. Néanmoins il faut pouvoir définir les facteurs de préférence d'une croyance, et le fonctionnement des biais cognitifs nous donne quelques pistes. En effet, le biais d'ancrage donne une importance plus grande à la première information reçue (ici $3 > 4$). Une notion de temps doit donc être ajoutée dans la logique afin de prendre en compte ce facteur. De plus, le temps n'est pas la seule notion importante si nous nous intéressons à d'autres biais. Par exemple le biais de confirmation qui est très proche de l'ancrage vient ignorer les informations qui rentrent en contradiction avec nos croyances de base. Comme le décrit Nickerson [24], une croyance ancienne A et appuyée par d'autres croyances B ($B \rightarrow A$) sera difficile à abandonner et sera sujette à un biais de confirmation. Là encore cette notion d'enracinement doit être prise en compte dans la sélection. Enfin les émotions jouent un rôle aussi dans les biais cognitifs. Par exemple le biais *attentional* [19], nous dirige vers des croyances qui ont un sens émotionnel. Par exemple 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). De plus, un biais n'est pas limité à un seul facteur mais peu aussi dépendre d'une combinaison de ceux-ci. Par exemple, [11] montre que le biais d'ancrage est plus prononcé lorsque nous sommes tristes.

Ainsi le travail futur reposera essentiellement sur la dé-

finition de fonctions de sélection correspondant à des biais spécifiques. Ces fonctions reposeront sur des facteurs d'évaluations avec notamment :

— le temps
— les émotions
— l'enracinement des croyances

De ce fait, ces fonctions permettront de mettre en évidence, dans une situation donnée, si un biais cognitif a pu jouer un rôle ou non. De plus nous voyons que plusieurs biais cognitifs prennent en compte une combinaison de tous ces facteurs et que nous devons trouver un moyen d'agencer tous ces éléments pour caractériser un biais. En s'appuyant sur la littérature des sciences humaines, une étude plus poussée des différents facteurs à prendre en compte dans le fonctionnement des biais cognitifs permettra sûrement de mettre en valeur d'autres facteurs mais aussi l'importance de ces facteurs. Par exemple nous pouvons voir que la notion de temps est centrale pour la conservation des croyances.

5 Conclusion

Dans cet article, nous avons proposé un modèle général pour diagnostiquer les biais cognitifs dans un modèle logique d'action, comme le modèle BDI [28] ou le calcul des situations [27]. Notre modèle a l'avantage de reposer sur la théorie de la révision de croyances qui a déjà été largement étudiée et formalisée dans la littérature. De plus, nous montrons que la génération des comportements possibles par la révision de croyances est suffisante pour retrouver des comportements similaires aux biais cognitifs tout en préservant la rationalité de la révision de croyances de l'agent. La prochaine étape de nos travaux est de caractériser la fonction de sélection venant choisir le comportement relatif à un biais spécifique. Nous avons donné dans cet article quelques premières pistes sur les facteurs nécessaires à intégrer pour déterminer un biais voulu. Ce dernier point fera l'objet d'un travail futur pour spécifier quels facteurs ont une influence sur quels types de biais. Nous pouvons pour ce point nous appuyer sur la littérature des sciences humaines. Pour l'intégration, la littérature de la révision de croyances étant déjà très fournie, de nombreux travaux se sont déjà penchés sur le problème du temps [6], des émotions [26] et de l'enracinement des croyances [13]. Enfin Arnaud et Adam [2] ont mis en évidence les limitations du modèle BDI pour rendre compte des comportements humains. Notamment dans l'engagement dans les intentions, la mise à jour des croyances et le changement de contexte. Les auteurs mettent en évidence l'importance de prendre en compte les sciences sociales et d'intégrer les biais cognitifs dans la BDI. Nous pensons que les prémisses de notre travail offrent un point de départ pour répondre à cette problématique.

Références

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

[2] Maël Arnaud, Carole Adam, and Julie Dugdale. Les limites du bdi pour rendre compte du comportement humain en situation de crise. 2017.

[3] Maël Arnaud, Carole Adam, and Julie Dugdale. The role of cognitive biases in reactions to bushfires. 2017.

[4] Buster Benson. Cognitive bias cheat sheet. *Better Humans*, 2016.

[5] Philippe Besnard, Éric Grégoire, and Jean-Marie Lagniez. On computing maximal subsets of clauses that must be satisfiable with possibly mutually-contradictory assumptive contexts. In *AAAI*, 2015.

[6] Giacomo Bonanno. Axiomatic characterization of the agm theory of belief revision in a temporal logic. *Artificial Intelligence*, 171(2-3) :144–160, 2007.

[7] Ori Brafman and Rom Brafman. *Sway : The irresistible pull of irrational behavior*. Crown Business, 2008.

[8] Marie-Odile Cordier, Philippe Dague, Yannick Pencolé, and Louise Travé-Massuyès. Diagnostic et supervision : approches à base de modèles. In Pierre Marquis, Odile Papini, and Henri Prade, editors, *Panorama de l'intelligence artificielle : Ses bases méthodologiques, ses développements*, volume 2. Cépaduès, January 2013.

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

[10] Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile. Bea f-cp090601. 2012.

[11] Birte Englich and Kirsten Soder. Moody experts—how mood and expertise influence judgmental anchoring. *Judgment and Decision making*, 4(1) :41, 2009.

[12] Eduardo L. Fermé and Sven Ove Hansson. Agm 25 years. *Journal of Philosophical Logic*, 40 :295–331, 2011.

[13] Peter Gärdenfors. Epistemic importance and minimal changes of belief. *Australasian Journal of Philosophy*, 62(2) :136–157, 1984.

[14] Peter Gärdenfors, Hans Rott, DM Gabbay, CJ Hogger, and JA Robinson. Belief revision. *Computational Complexity*, 63(6), 1995.

[15] Andrew Gilbey and Stephen Hill. Confirmation bias in general aviation lost procedures. *Applied Cognitive Psychology*, 26(5) :785–795, 2012.

[16] Julien Goldszlagier. L'effet d'ancrage ou l'apport de la psychologie cognitive à l'étude de la décision judiciaire. *Les Cahiers de la Justice*, (4) :507–531, 2015.

[17] Sven Ove Hansson. Kernel contraction. *The Journal of Symbolic Logic*, 59(3) :845–859, 1994.

[18] Jonathan Kulick and Paul K Davis. Modeling adversaries and related cognitive biases. 2003.

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

[20] Shem Malmquist. The role of cognitive bias in aircraft accident. 2014.

[21] Joao Marques-Silva, Federico Heras, Mikolás Janota, Alessandro Previti, and Anton Belov. On computing minimal correction subsets. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

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

[23] Clifford R Mynatt, Michael E Doherty, and Ryan D Tweney. Confirmation bias in a simulated research environment : An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29(1) :85–95, 1977.

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

[25] ED O'Sullivan and SJ Schofield. Cognitive bias in clinical medicine. *JR Coll Physicians Edinb*, 48(3) :225–32, 2018.

[26] César F Pimentel and Maria R Cravo. Affective revision. In *Portuguese Conference on Artificial Intelligence*, pages 115–126. Springer, 2005.

[27] Javier Andres Pinto and Raymond Reiter. *Temporal reasoning in the situation calculus.* University of Toronto, 1994.

[28] Anand S Rao, Michael P Georgeff, et al. Bdi agents : from theory to practice. In *ICMAS*, volume 95, pages 312–319, 1995.

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

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

[31] Peter M Todd and Gerd Gigerenzer. Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, 23(5) :727–741, 2000.

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

[33] Amos Tversky and Daniel Kahneman. Judgment under uncertainty : Heuristics and biases. *science*, 185(4157) :1124–1131, 1974.

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

[35] Stephen Walmsley and Andrew Gilbey. Cognitive biases in visual pilots' weather-related decision making. *Applied Cognitive Psychology*, 30(4) :532–543, 2016.

[36] Renata Wassermann. An algorithm for belief revision. 2000.